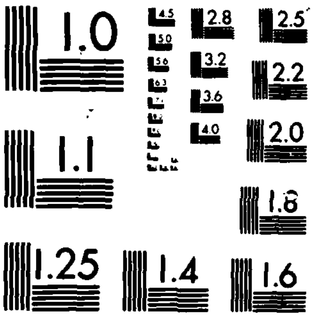
AD-A164 368 THE FACTORIZATION METHOD FOR THE NUMERICAL SOLUTION OF 1/1
TWO POINT BOUNDARY..(U) MARYLAND UNIV COLLEGE PARK INST
FOR PHYSICAL SCIENCE AND TECH.. I BABUSKA ET AL.
UNCLASSIFIED JAN 86 BN-1039 N00014-85-K-0169 F/G 12/1 NL

END

FILMED



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

15

INSTITUTE FOR PHYSICAL SCIENCE
AND TECHNOLOGY

Laboratory for Numerical Analysis

Technical Note BN-1039

THE FACTORIZATION METHOD FOR THE NUMERICAL SOLUTION OF TWO POINT BOUNDARY VALUE PROBLEMS FOR LINEAR ODE'S

by

I. Babuška and V. Majer
Institute for Physical Science and Technology
University of Maryland, College Park, 20742

January 1986

UNIVERSITY OF MARYLAND

Encl (1)

REPORT DOCUMENTATION PAGE		**READ INSTRUCTIONS BEFORE COMPLETING FORM**
1. REPORT NUMBER Technical Note BN-1039	2. GOVT ACCESSION NO. ADA164368	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Factorization Method for the Numerical Solution of Two Point Boundary Value Problems for Linear ODE's		5. TYPE OF REPORT & PERIOD COVERED Final life of the contract 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) I. Babuška and V. Majer		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-85-K-0169
9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Physical Science and Technology University of Maryland College Park, Maryland 20742		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of Naval Research Office of Naval Research Arlington, VA 22217		12. REPORT DATE January, 1986 13. NUMBER OF PAGES 74
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The theoretical analysis and computational implementation of factorization-based methods for the numerical solution of linear boundary value problems for ordinary differential equations are presented. The methods are optimal with respect to certain clearly defined criteria. Numerical examples show the effectiveness of a general code based on the factorization method.

THE FACTORIZATION METHOD FOR THE NUMERICAL SOLUTION OF TWO POINT BOUNDARY VALUE PROBLEMS FOR LINEAR ODE'S

by

I. Babuška* and V. Majer**

*Partially supported by the ONR-N-00014-85-K-0169.

**Supported by NSF Grant DMS-8315216.

ABSTRACT

The theoretical analysis and computational implementation of factorization-based methods for the numerical solution of linear boundary value problems for ordinary differential equations are presented. The methods are optimal with respect to certain clearly defined criteria. Numerical examples show the effectiveness of a general code based on the factorization method.

1 INTRODUCTION

1.1 Numerical methods for linear boundary value problems for ordinary differential equations

The standard techniques for the numerical solution of BVP's for ODE's can be divided into two classes. On the one hand are the "direct" methods based on various versions of finite differences, finite elements, or collocation, and on the other are the "indirect" or sequential methods based on the numerical solution of auxiliary initial value problems. Typical of this class are various shooting and multi-shooting approaches.

Direct methods are characterized by the solution of global (linear) algebraic systems for the discrete solution. (In this sense multi-shooting may be regarded as a hybrid method between the two classes. A sophisticated recent example of multi-shooting is the BOUNDPAC package of Mattheij and Staarink [17]) The solution of the linear systems, which should of course be regarded as part of the numerical solution of the original problem, can be accomplished by an iterative method or by a direct method based on elimination. The discretization itself may be adaptive or non-adaptive; adaptivity, however, generally requires multiple solutions of the problem. The well-known programs COLSYS [1] and PASVA [16] are based on methods from this class.

Indirect solution methods are characterized by the association of the BVP with certain auxiliary initial value problems (IVP). The auxiliary IVP's are generally solved uni-directionally (forward), but a subclass of initial value based methods are based on bi-directional (double sweep) strategies. We use the term factorization for this subclass, and it is to the analysis and exploitation of the underlying structure of the methods of this class that this paper is addressed.

Both classes mentioned above are well-represented in the literature. In [2], [3], [13] and [15] the relationship between methods of the two classes is explored. The connection is made through the notion of the closure of a numerical algorithm ([5], [6]). It can be shown that if one incorporates the numerical solution of the discretized (algebraic) equations into the algorithm for direct methods, then the algorithm can be interpreted as the application of a special sequential numerical solver for some naturally associated initial value problems. We refer here to [3], [11], [13], [15] and [18] for some recent papers addressing in various ways this relationship. In this paper we explore the subclass of indirect methods called factorizations, discuss the principles of their adaptive construction by the computer itself, and address certain questions of implementation. Numerical examples illustrate the effectiveness of a general code based on the method.

1.2 The BVP and the goals of the computations

Consider the linear two point boundary value problem

$$w'(s) = B(s)w(s) - F(s), \qquad s_1 \leq s \leq s_2, \tag{1a}$$

$$U_1 w(s_1) = u_1, \qquad U_2 w(s_2) = u_2, \tag{1b}$$

where $B(\cdot)$ is an $n \times n$ matrix function, $w(\cdot)$ and $F(\cdot)$ are n-vector functions, U_1 and U_2 are $n_1 \times n$ and $n_2 \times n$ matrices, respectively, and u_1 and u_2 are and n_1- and n_2-vectors, respectively.

The goals of the computations are as follows: given a set of target points

$$s_1 = \sigma_0 < \sigma_1 < \cdots < \sigma_n = s_2, \tag{2}$$

and a tolerance, τ, find vectors w_i such that

(3) $$w_i = \hat{w}_i(\sigma_i), \qquad i = 0,\ldots,n$$

where $\hat{w}_i(\cdot)$ is the exact solution of a perturbed problem

(4a) $$\hat{w}_i'(s) = (B(s) + b_i(s))\hat{w}_i(s) - (F(s) + f_i(s)), \qquad s_1 \le s \le s_2,$$

(4b) $$(U_\alpha + V_\alpha)\hat{w}_i(s_\alpha) = (u_\alpha + v_\alpha), \qquad \alpha = 1,2,$$

where the perturbations may depend on the target point, as indicated, but satisfy

(5) $$|b_i(\cdot)|, \; |f_i(\cdot)|, \; |V_\alpha|, \; |v_\alpha| \le \tau.$$

The norms in (5) are a-priori selected and may, in the case of b_i and f_i, be of the type L_p, $1 \le p \le \infty$. The vector w_i is a trace of an exact solution of the problem (1) with perturbed input data; the perturbations depend on σ_i and may be different for different i.

Obviously the aim of the computation is directly related to the interpretation of the numerical solution of engineering problems where the input data are not known precisely. That is, the class of perturbed problems (4) are, for perturbations of a known magnitude τ, indistinguishable with respect to the engineering interpretaion of their exact solutions.

1.3 Factorization methods and their adaptive construction

We consider a class of methods based on the bi-directional adaptive solution of IVP's for certain ODE's which themselves are selected adaptively. If these equations were solved exactly then the exact solution of (1) would be obtained at each target point. Let us assume that, in the forward direction, say, these IVP's are of the form

$$\Phi'(s) = R(s,\Phi(s)), \qquad s_1 \le s, \tag{6a}$$

$$\Phi(s_1) = \Phi_1. \tag{6b}$$

We cannot obtain Φ exactly, but only an approximate solution $\tilde{\Phi}$ which solves (exactly) the equation

$$\tilde{\Phi}'(s) = R(s,\tilde{\Phi}(s)) + r(s) \tag{7}$$

where $|r(\cdot)| \le \tau$, the given solution tolerance, and $|\cdot|$ is a suitable L_p norm. In [7] it is shown how various local error control strategies achieve $|r| \le \tau$ for different L_p norms using a minimal number of steps.

We study a class of methods which directly ties the norm $|r|$ to the perturbations of the input data of the original BVP in the sense that $|b_i|$, $|f_i|$ $|V_\alpha|, |v_\alpha| \le C|r|$ where C is an a-priori known constant independent of the problem (1) with $C = 1$. Not only are the ODE's (7) solved adaptively in order to ensure the tolerance τ, but the ODE's themselves are constructed adaptively in order to ensure that $C = 1$. It is in this precise sense that the methods we consider are optimal.

1.4 Outline of the paper

In Section 2 we formulate and analyze the class of methods which satisfy the requirements stated above. Section 3 focuses on the sub-class of factorizations which reduce to matrix Riccati equations. We derive and analyze a special solver for such matrix Riccati equations. The Riccati solver is the foundation of a general linear TPBVP code whose performance is illustrated on a varied selection of example problems in Section 4.

The efficient and cost-effective solution of linear problems has enabled this method to be applied also to large-scale nonlinear BVP's with turning points and bifurcations. The description of this approach, its analysis, and our computational experience will be reported elsewhere.

2 THE FACTORIZATION METHOD

2.1 The linear two-point boundary value problem

Definition 1.

(a) Matrix B is said to be of size (m,n) if it has m rows and n columns.

(b) Matrix B of size (m,n) is said to be of type (m,n) if it has maximal rank. ■

Suppose that bounded and measurable matrix functions $s \to B(s)$ of size (m,n) and $s \to F(s)$ of size $(n,1)$, $n > 0$, are given on the interval $[s_1,s_2] \subset R^1$. Suppose also that for $\alpha = 1,2$ matrices U_α of type (n_α,n) and u_α of size $(n_\alpha,1)$ with $n_1 + n_2 = n$ are given. We seek an absolutely continuous (a.c.) vector function $s \to w(s)$ of size $(n,1)$ on $[s_1,s_2]$ which solves the first order linear two point boundary value problem (TPBVP)

$$(1) \qquad \begin{cases} w'(s) = B(s)w(s) - F(s), & \text{a.e. } s \in [s_1,s_2], \\ U_1 w(s_1) = u_1, \quad U_2 w(s_2) = u_2. \end{cases}$$

We will refer to the TPBVP (1) as $P(B,F,U_\alpha,u_\alpha,s_\alpha)$.

The separated boundary conditions in (1) are no real restriction, for the mixed problem

$$(2) \qquad \begin{cases} w'(s) = B(s)w(s) - F(s), & \text{a.e. } s \in [s_1,s_2] \\ U_1 w(s_1) + U_2 w(s_2) = u \end{cases}$$

can easily be cast into the form (1). The algorithms for the solution of P that we consider are based on the integration of certain associated initial value problems. Accordingly, we turn to the development of some auxiliary results concerning matrix initial value problems.

2.2 Auxiliary results for matrix initial value problems

Lemma 1. Let matrix functions $s \to B(s), F(s)$ of size (n,n) and size $(n,1)$, respectively, be given on $[s_1, s_2]$. Suppose that matrix U is of type (p,n), $p \leq n$ and that u is of size $(p,1)$. Suppose further that $s \to v(s)$ of size $(n,1)$ is absolutely continuous and satisfies

$$(1) \quad \begin{cases} v'(s) = B(s)v(s) - F(s) & \text{a.e. } s \in [\sigma_1, r_2] \subset [s_1, s_2] \\ Uv(\sigma_0) = u & \text{for some } \sigma_0 \in [\sigma_1, \sigma_2]. \end{cases}$$

If $s \to \Phi(s)$ of size (p,n) defined on $[\sigma_1, \sigma_2]$ is a.c. and satisfies

$$(2) \quad \begin{cases} \Phi'(s) = -\Phi(s)B(s) + Z(s)\Phi(s), & \text{a.e. } s \in [\sigma_1, \sigma_2] \\ \Phi(\sigma_0) = U, \end{cases}$$

and if $s \to \varphi(s)$ of size $(p,1)$ is a.c. and satisfies

$$(3) \quad \begin{cases} \varphi'(s) = -\Phi(s)F(s) + Z(s)\varphi(s), & \text{a e. } s \in [\sigma_1, \sigma_2] \\ \varphi(\sigma_0 = u, \end{cases}$$

where $s \to Z(s)$ of size (p,p) on $[\sigma_1, \sigma_2]$ is continuous, but otherwise arbitrary, then,

$$(4) \qquad \Phi(s)v(s) = \varphi(s) \qquad \forall s \in [\sigma_1, \sigma_2]$$

Proof. Define $\psi(s)$ on $[\sigma_1, \sigma_2]$ by

$$\psi(s) = (\Phi v - \varphi)(s).$$

Then

$$\psi(\sigma_0) = Uv(\sigma_0) - u = 0,$$

and

$$\begin{aligned}\psi' &= \Phi' v + \Phi v' - \varphi' \\ &= (-\Phi B + Z\Phi)v + \Phi(Bv - F) - (\Phi F + Z\ ) \\ &= Z(\Phi v - \varphi) = Z\psi, \qquad \text{a.e. } s \in [\sigma_1, \sigma_2].\end{aligned}$$

Since Z is continuous, $\psi \equiv 0$ by uniqueness and the lemma is proved. ∎

The method of factorization for BVP's is based on the propagation of the boundary condition across the interval in a manner consistent with the differential equation (1); Lemma 1 is the formal expression of the nature of this propagation. We use the following terminology.

Definition 1. A matrix function Φ of size (p,n) satisfying (2) is a transition matrix based at σ_0. The matrix Z which induces Φ is the associated conditioning matrix on $[\sigma_1, \sigma_2] \subset [s_1, s_2]$. The vector function φ of size $(p,1)$ satisfying (3) is a transition vector based at σ_0. ∎

Lemma 2. Suppose that Φ and Ψ are size (p,n) transition matrices on $[\sigma_1, \sigma_2]$ based at σ_0, Z and Y are continuous size (p,p) conditioning matrices on $[\sigma_1, \sigma_2]$, that

(5) $$\Phi' = -\Phi B + Z\Phi,$$

(6) $$\Psi' = -\Psi B + Y\Psi,$$

on $[\sigma_1,\sigma_2]$, and that

$$\Phi(\sigma_0) = K_0\Psi(\sigma_0) \tag{7}$$

with K_0 of type (p,p). Then there exists a matrix function $s \to K(s)$ of type (p,p) on $[\sigma_1,\sigma_2]$ such that,

$$\Phi(s) = K(s)\,\Psi(s), \qquad s \in [\sigma_1,\sigma_2]. \tag{8}$$

Proof. Let $s \to K(s)$ solve the (linear) initial value problem

$$\begin{cases} K'(s\ ) = Z(s)K(s) - K(s)Y(s), \\ K(\sigma_0) = K_0. \end{cases} \tag{9}$$

Then we have that on $[\sigma_1,\sigma_2]$

$$(K\Psi-\Phi)' = -(K\Psi-\Phi)B + Z(K\Psi-\Phi) \tag{10}$$

with initial condition

$$(K\Psi-\Phi)(\sigma_0) = 0. \tag{11}$$

By uniqueness of solutions to (10), (11) we have that

$$\Phi(s) = K(s)\Psi(s), \qquad s \in [\sigma_1,\sigma_2]. \tag{12}$$

It remains to show that K is of type (p,p), i.e., invertible. Let $L(s)$ be the solution of

$$\begin{cases} L'(s) = Y(s)L(s) - L(s)Z(s), \\ L(\sigma_0) = K_0^{-1} \end{cases}$$

then KL satisfies

$$(KL)' = Z(KL) - (KL)Z \tag{13}$$

$$(KL)(\sigma_0) = Id_p. \tag{14}$$

Since the $p \times p$ identity matrix Id_p is the unique solution of (13), (14) on $[\sigma_1, \sigma_2]$, the result is proved. ■

Lemma 3. The rank of $\Phi(s)$ satisfying (5) is constant on $[\sigma_1, \sigma_2]$.

Proof. Let $s \to E(s)$ of size (n,n) be the solution of

$$\begin{cases} E'(s) = -E(s)B(s) & s \in [\sigma_1, \sigma_2] \\ E(\sigma_0) = Id_n. \end{cases}$$

Clearly Ψ solving (6) with $Y = 0$ exists and is given by

$$\Psi(s) = \Psi(\sigma_0)E(s),$$

so that by Lemma 2

$$\Phi(s) = K(s)\Psi(\sigma_0)E(s). \tag{15}$$

If we show that E is non-singular for $s \in [\sigma_1, \sigma_2]$ then from (15) we can see that

$$\operatorname{rank} \Phi(s) = \operatorname{rank} \Psi(s)$$

which proves the result.

To show that $E(s)$ is of type (n,n), let $F(s)$ solve

$$\begin{cases} F'(s) = B(s)F(s) \\ F(\sigma_0) = Id_n. \end{cases} \tag{16}$$

It is easy to see that $E(s)F(s) = Id_n$ on $[\sigma_1, \sigma_2]$. ■

We apply Lemma 3 to show that solvability of (4) at one point $\sigma_0 \in [\sigma_1, \sigma_2]$ gives solvability at each $s \in [\sigma_1, \sigma_2]$. In fact we can prove

Corollary 4. For Φ, φ satisfying (2), (3), if the equation

$$Uw_0 = u$$

has k independent solutions, then

$$\Phi(s)w = \varphi(s)$$

has k independent solutions for every $s \in [\sigma_1, \sigma_2]$.

Proof. Apply Lemmas 1 and 3 to the augmented matrix $[\Phi \mid \varphi]$ which solves

$$[\Phi \mid \varphi] = -[\Phi \mid \varphi] \begin{bmatrix} B & F \\ 0 & 0 \end{bmatrix},$$

(17)

$$[\Phi \mid \varphi](\sigma_0 = [U \mid u].$$

By by hypothesis, rank $U = n - k = p$ and $u \in$ range U. Now rank$[\Phi \ \varphi] = p$ and rank $\Phi = p$ by Lemma 3. Thus $\varphi \in$ range Φ and the proof is complete. ■

2.3 Definition of a factorization

Let boundary value problem $\mathcal{P}(B, F, U_\alpha, u_\alpha, s_\alpha)$ be given as in Section 2.1

Definition 1. A <u>factorization</u> of $\mathcal{P}$ consists of:

(Fa) partitions π_α, $\alpha = 1,2$ of $[s_1,s_2]$ with

$$\pi_1 : s_1 = \sigma_1^{(0)} < \sigma_1^{(1)} < \cdots < \sigma_1^{(m_1)} = s_2,$$

$$\pi_2 : S_1 = \sigma_2^{(m_2)} < \sigma_2^{(m_2-1)} < \cdots < \sigma_2^{(0)} = s_2;$$

(Fb) collections Z_α, $\alpha = 1,2$ of conditioning matrices of size (n_α,n_α),

$$Z_\alpha = \{s \to Z_\alpha^{(i)}(s) : s \in I_\alpha^{(i)}, \quad i = 1,m_\alpha\},$$

where we use the notation

$$I_\alpha^{(i)} = \begin{cases} [\sigma_1^{(i-1)},\sigma_1^{(i)}], & \alpha = 1, \\ [\sigma_2^{(i)},\sigma_2^{(i-1)}], & \alpha = 2; \end{cases}$$

(Fc) collections Φ_α, $\alpha = 1,2$ of transition matrices of type (n_α,n),

$$\Phi_\alpha = \{s \to \Phi_\alpha^{(i)}(s) : s \to I_\alpha^{(i)}, \quad i = 1,m_\alpha\};$$

(Fd) collections φ_α, $\alpha = 1,2$ of transition vectors of size $(n_\alpha,1)$,

$$\varphi_\alpha = \{s \to \varphi_\alpha^{(i)}(s) : s \in I_\alpha^{(i)}, \quad i = 1,m_\alpha\};$$

(Fe) collections K_α, $\alpha = 1,2$ of *scaling matrices* of type (n_α,n_α),

$$K_\alpha = \{K_\alpha^{(i)}, \quad i = 0, m_\alpha - 1\};$$

and

(Ff) collections P_α, $\alpha = 1,2$ of constant <u>similarity</u> <u>matrices</u> of type (n,n),

$$P_\alpha = \{P_\alpha^{(i)},\ i = 1,m_\alpha\},$$

such that the following conditions hold for $i = 1,m_\alpha$ and $\alpha = 1,2$.

(F1) $$B_\alpha^{(i)}(s) = (P_\alpha^{(i)})^{-1}B(s)P_\alpha^{(i)}, \qquad s \in I_\alpha^{(i)};$$

(F2) $$F_\alpha^{(i)}(s) = (P_\alpha^{(i)})^{-1}F(s),\ s \quad I_\alpha^{(i)};$$

(F3) $$\Phi_\alpha^{(i)\prime}(s) = -\Phi_\alpha^{(i)}(s)B_\alpha^{(i)}(s) + Z_\alpha^{(i)}(s)\Phi_\alpha^{(i)}(s), \qquad s \in I_\alpha^{(i)};$$

(F4) $$\varphi_\alpha^{(i)\prime}(s) = -\Phi_\alpha^{(i)}(s)F_\alpha^{(i)}(s) + Z_\alpha^{(i)}(s)\varphi_\alpha^{(i)}(s), \qquad s \in I_\alpha^{(i)};$$

(F5) $$\Phi_\alpha^{(i)}(\sigma_\alpha^{(i-1)}) = K_\alpha^{(i-1)}\Phi_\alpha^{(i-1)}(\sigma_\alpha^{(i-1)})P_\alpha^{(i)}$$

(F6) $$\varphi_\alpha^{(i)}(\sigma_\alpha^{(i-1)}) = K_\alpha^{(i-1)}\varphi_\alpha^{(i-1)}(\sigma_\alpha^{(i-1)}).$$

In order for (F5) and (F6) to make sense for $i = 1$, we have used the notational conventions

(F7) $$\Phi_\alpha^{(0)}(\sigma_\alpha^{(0)}) = U_\alpha,$$

and

(F8) $$\varphi_\alpha^{(0)}(\sigma_\alpha^{(0)}) = u_\alpha.$$

The initial value problems (F3), (F5) and (F4), (F6) for $\Phi_\alpha^{(i)}$

and $\varphi_{\alpha}^{(i)}$, respectively, are posed forward in s for $\alpha = 1$ and backward in s for $\alpha = 2$. We combine the forward and backward factorizations into a composite factorization as follows:

Let the composite partition $\pi = \pi_1 \cup \pi_2$ be given by

$$\pi : s_1 = \sigma^{(0)} < \sigma^{(1)} < \cdots < \sigma^{(m)} = s_2, \tag{1}$$

and let $I^{(i)} = [\sigma^{(i-1)}, \sigma^{(i)}]$, $i = 1,m$. It is obvious that

Lemma 1. There exist unique indices i_{α}, $\alpha = 1,2$ such that

$$I^{(i)} = I^{(i_2)} \cap I^{(i_2)}. \qquad \blacksquare$$

For $\sigma \in I^{(i)} = I_1^{(i_1)} \cap I_2^{(i_2)}$ we define $\Phi^{(i)}(\sigma)$ of size (n,n) by

$$\Phi_{\sigma}^{(i)} = \begin{bmatrix} \Phi_1^{(i_1)}(\sigma)(P_1^{(i_1)})^{-1} \\ \\ \Phi_2^{(i_2)}(\sigma)(P_2^{(i_2)})^{-1} \end{bmatrix}, \tag{2}$$

$\varphi^{(i)}(\sigma)$ of size $(n,1)$ by

$$\varphi^{(i)}(\sigma) = \begin{bmatrix} \varphi_1^{(i_1)}(\sigma) \\ \\ \varphi_2^{(i_2)}(\sigma) \end{bmatrix}, \tag{3}$$

and

(4) $$Z^{(i)}(\sigma) = \left[\begin{array}{c|c} Z_1^{(i_1)}(\sigma) & 0_{n_1,n_2} \\ \hline 0_{n_2,n_1} & Z_2^{(i_2)}(\sigma) \end{array}\right].$$

The definition of the composite scaling matrices is more involved. For $i = 0,\dots,m$ define

(5a) $$L_1^{(i)} = \begin{cases} K_1^{(i_1)} & \text{if } \sigma^{(i)} = \sigma^{(i_1)}, \\ Id_{n_1} & \text{otherwise,} \end{cases}$$

(5b) $$L_2^{(i)} = Id_{n_2},$$

and then set

(5c) $$L^{(i)} = \left[\begin{array}{c|c} L_1^{(i_1)} & 0_{n_1,n_2} \\ \hline 0_{n_2,n_1} & L_2^{(i)} \end{array}\right].$$

Similarly, define

(6a) $$R_1^{(i)} = Id_{n_1},$$

(6b) $$R_2^{(i)} = \begin{cases} K_2^{(i_2)} & \text{if } \sigma^{(i)} = \sigma^{(i_2)} \\ Id_{n_2} & \text{otherwise} \end{cases}$$

and then set

$$\text{(6c)} \qquad R^{(i)} = \left[\begin{array}{c|c} R_1^{(i)} & 0_{n_1,n_2} \\ \hline 0_{n_2,n_1} & R_2^{(i)} \end{array}\right].$$

Finally, we set

$$\text{(7)} \qquad K^{(i)} = (R^{(i)})^{-1}L^{(i)}$$

and then state

Lemma 2. $\Phi^{(i)}$ is of size (n,n), $\varphi^{(i)}$ is of size $(n,1)$, $Z^{(i)}$ is of size (n,n), $L^{(i)}$ is of type (n,n), $R^{(i)}$ is of type (n,n) (and so $K^{(i)}$ is well-defined), and we have

$$\text{(8)} \qquad \Phi^{(i)'} = -\Phi^{(i)}B + Z^{(i)}\Phi^{(i)} \quad \text{a.e. on } I^{(i)}, \quad i = 1,\ldots,m,$$

$$\text{(9)} \qquad \varphi^{(i)'} = -\Phi^{(i)}F + Z^{(i)}\Phi^{(i)} \quad \text{a.e. on } I^{(i)}, \quad i = 1,\ldots,m,$$

$$\text{(10)} \qquad \Phi^{(i)}(\sigma^{i-1}) = K^{i-1}\Phi^{(i-1)}(\sigma^{(i-1)}), \qquad i = 1,\ldots,m+1$$

$$\text{(11)} \qquad \varphi^{(i)}(\sigma^{i-1}) = K^{i-1}\varphi^{(i-1)}(\sigma^{(i-1)}), \qquad i = 1,\ldots,m+1.$$

Proof. Equations (8)-(11) follow directly from (F1)-(F6).

We also establish a notation for the composite factorization.

Definition 2. A composite factorization, $\mathcal{F}$ for $\mathcal{P}(B,F,U_\alpha,u_\alpha,s_\alpha)$ consists of the sets

$$\Phi = \{s \to \Phi^{(i)}(s) : s \quad I^{(i)}, \quad \Phi^{(i)} \text{ satisfies (8)}\},$$

$$\varphi = \{s \to \varphi^{(i)}(s) : s \in I^{(i)}, \quad \varphi^{(i)} \text{ satisfies (10)}\},$$

$$Z = \{s \to Z^{(i)}(s) : s \quad I^{(i)}, \quad Z^{(i)} \text{ as in (4)}\},$$

$$K = \{K^{(i)} : K^{(i)} \text{ as in (7)}\}.$$

We write $F = F(\pi,\Phi,\varphi,Z,K) \equiv \{\pi,\Phi,\varphi,Z,K\}$. ■

Theorem 1. Suppose $F(\pi,\Phi,\varphi,Z,K)$ is a factorization for $P(B,F,U_\alpha,u_\alpha,s_\alpha)$ and let $s \to w(s)$ be an a.c. solution of P. Then

$$\Phi w = \varphi. \tag{12}$$

The sense in which (11) is to be interpreted will be apparent from the

Proof. Let a <u>target point</u> $\sigma \in [s_1,s_2]$ be given. Then $\sigma \in I^{(i)}$ for some index i, $0 \le i \le m$. We must show that

$$\Phi^{(i)}(\sigma)w(\sigma) = \varphi^{(i)}(\sigma). \tag{13}$$

Since $I^{(i)} = I_1^{(i_1)} \cap I_2^{(i_2)}$ it is enough to show that

$$\Phi_\alpha^{(i_\alpha)}(s)(P_\sigma^{(i_\alpha)})^{-1} w(s) = \varphi_\alpha^{(i_\alpha)}(s)\ \forall s \in I_\alpha^{(i_\alpha)}, \quad \alpha = 1,2. \tag{14}$$

We prove (14) by induction on i_α.

<u>$i_\alpha = 1$:</u> By definition of P we have that

$$U_\alpha w(\sigma_\alpha^{(0)}) = u_\alpha$$

and so

$$K_\alpha^{(0)} U_\alpha w(\sigma_\alpha^{(0)}) = K_\alpha^{(0)} u_\alpha .$$

But by (F5) and (F7) we have

$$\Phi_\alpha^{(1)}(\sigma^{(0)}) = K_\alpha^{(0)} U_\alpha P_\alpha^{(1)}$$

and by (F6) and (F8)

$$\varphi_\alpha^{(1)}(\sigma_\alpha^{(0)}) = K_\alpha^{(0)} u_\alpha .$$

Now by virtue of Lemma 2.2-1, (14) holds for $i_\alpha = 1$. In particular, (14) holds for $s = \sigma_\alpha^{(1)}$ and $i_\alpha = 1$. Since $K_\alpha^{(1)}$ is of type (n_α, n_α), this gives

$$K_\alpha^{(1)} \Phi_\alpha^{(1)} (P_\alpha^{(1)})^{-1} w(\sigma_\alpha^{(1)}) = K_\alpha^{(1)} \varphi(\sigma_\alpha^{(1)}). \tag{15}$$

<u>i_α to $i_\alpha + 1$:</u> Having

$$K_\alpha^{(i_\alpha)} \Phi_\alpha^{(i_\alpha)} (P_\alpha^{(i_\alpha)})^{-1} w(\sigma_\alpha^{(i_\alpha)}) = K_\alpha^{(i_\alpha)} \varphi(\sigma_\alpha^{(i_\alpha)})$$

and (F3)-(F6) for $i = i_\alpha + 1$ we again apply Lemma 2.2-1 to conclude that (14) holds for $i = i_\alpha + 1$. ■

The proof of Theorem 1 actually gives us a bit more than is stated in the theorem. The induction together with the observation that each $K_\alpha^{(i_a)}$ is of type (n_α, n_α), i.e., invertible, and an appeal to Lemma 2.2-3 gives

Corollary 2. The rank of Φ_α, $\alpha = 1,2$, is constant on $[s_1,s_2]$. That is, Φ_α is of type (n_α,n) on $[s_1,s_2]$. ■

Now that we have shown that (11) holds if w solves P, we show that the solvability of (11) is also inherited from P. In fact, they are equivalent.

Theorem 3. Let $F(\pi,\Phi,\varphi,Z,K)$ be a factorization for $P(B,F,U_\alpha,u_\alpha,s_\alpha)$. Then for $\sigma \in [s_1,s_2]$, the linear algebraic equation

$$\Phi(\sigma)w = \varphi(\sigma) \tag{15}$$

has as many linearly independent solutions $w \in \mathbf{R}^n$ as there are linearly independent a.c. functions $s \to w(s)$ solving P.

Proof. Equation (15) is interpreted in the sense of Theorem 1.

If $s \to w(s)$ solves P then by Theorem 1 $w = w(\sigma)$ is a solution of (15).

On the other hand, let w_1 be a solution of

$$\Phi(s_1)w_1 = \varphi(s_1)$$

and let $s \to w(s)$ be the solution of the initial value problem

$$\begin{cases} w'(s) = B(s)w(s) - F(s), s > s_1 \\ w(s_1) = w_1. \end{cases}$$

By the definition of Φ_1 and φ_1 and by the fact that $K_1^{(0)}$ is of type (n_1,n_1) we have

$$U_1 w(s_1) = u_1.$$

We must show that

$$U_2 w(s_2) = u_2. \tag{16}$$

Now by Lemma 2.2-1 applied with $p = n = n_1 + n_2$, $U = \Phi(s_1)$, and $u = \varphi(s_1)$, we conclude that

$$\Phi(s_2)w(s_2) = \varphi(s_2)$$

from which

$$\Phi_2(s_2)w(s_2) = \varphi_2(s_2).$$

Therefore

$$K_2^{(0)}U_2w(s_2) = K_2^{(0)}u_2$$

and since $K_2^{(0)}$ us of type (n_2,n_2) we have (16). ∎

In our development above we have assumed that a factorization for $\mathcal{P}$ exists. In fact, there are many, as we will show by example below. The trivial factorization $(Z = 0,\ K = Id_n)$ always exists, for example; it is usually not numerically realizable, for it is the shooting method. A factorization algorithm, then, should select the conditioning matrices $Z_\alpha^{(i)}$, the scaling matrices $K_\alpha^{(i)}$, the similarity matrices $P_\alpha^{(i)}$, and the partitions π_α adaptively in order to ensure the numerical stability of the computations.

2.4 Stable factorizations

From the point of view of practical computations, the boundary value problem $\mathcal{P}(B,F,U_\alpha,u_\alpha,s_\alpha)$ is but a model of physical reality. The data--the arguments of $\mathcal{P}$--are by definition known inaccurately, and it is only after the influences of these inaccuracies on the solution, w, to $\mathcal{P}$ are acknowledged that a reasonable interpretation of the meaning of even the exact solution to $\mathcal{P}$ can be made. This interpretation is based on our confidence both in the model itself and on the reliability of the data supplied. We take the point of view that the exact solutions w and $\hat{w}$ of $\mathcal{P}$ and a perturbed problem $\hat{\mathcal{P}}$, respectively, are equivalent if $\mathcal{P}$ and $\hat{\mathcal{P}}$ are indistinguishable with respect to the goals of the computation.

On the other hand, we do not have at our disposal the exact solution w to $\mathcal{P}$, but only an approximate solution, $\hat{w}$. However, if it is possible to interpret $\hat{w}$ as the exact solution of a perturbed problem $\hat{\mathcal{P}}$, and if $\mathcal{P}$ and $\hat{\mathcal{P}}$ are acceptably close (with respect to the goals of the computation), then $\hat{w}$ is an acceptable approximation to w.

In order to formalize this idea, let us suppose that a norm $|\cdot|_n$ is given on R^n. Then $|\cdot|_n$ induces a natural norm $|\cdot|_{(n,n)}$ on matrices of size (n,n):

$$(1) \qquad |A|_{(n,n)} = \sup_{|v|_n=1} |Av|_n .$$

We will need to measure the size of various sub-matrices of A in a consistent way. Suppose that A is a matrix of size (p,q), $1 \le p,q \le n$. Then

$$(2) \qquad |A|_{(p,q)} = |\tilde{A}|_{(n,n)}$$

where $\tilde{A}$ is obtained from A by augmentation by zero:

$$\tilde{A} = \begin{array}{cl} \left[\begin{array}{c|c} A & 0 \\ \hline 0 & 0 \end{array}\right] & \begin{array}{l} p \\ n-p \end{array} \\ \begin{array}{cc} q & n-q \end{array} & \end{array} \tag{3}$$

Of course, $|v|_n = |v|_{(n,1)} = |v^T|_{(1,n)}$ for $v \in \mathbb{R}^n$.

Now we are in a position to state

Definition 1. An approximate solution $\hat{w}$ of $P(B,F,U_\alpha u_\alpha,s_\alpha)$ is <u>δ-acceptable</u> at $\sigma \in [s_1,s_2]$ if $\hat{w}(\sigma)$ is the exact solution of a perturbed problem $\hat{\mathcal{P}} = \mathcal{P}(\hat{B},\hat{F},\hat{U}_\alpha,\hat{u}_\alpha,s_\alpha)$ where

$$|B(\cdot) - \hat{B}(\cdot)|_{(n,n)} \le \delta_B, \tag{4a}$$

$$|F(\cdot) - \hat{F}(\cdot)|_{(n,n)} \le \delta_F, \tag{4b}$$

$$|U_\alpha - \hat{U}_\alpha|_{(n_\alpha,1)} \le \delta_{U_\alpha}, \ \alpha = 1,2, \tag{4c}$$

$$|u_\alpha - \hat{u}_\alpha|_{(n_\alpha,n)} \le \delta_{u_\alpha}, \ \alpha = 1,2, \tag{4d}$$

and the norms (4a) and (4b) are understood as norms in L_p, $1 \le p \le \infty$. We use the notation $\delta = \{\delta_B,\delta_F,\delta_{U_\alpha},\delta_{u_a}\}$.

Based on the notion of δ-acceptability of the approximate solution, we seek to formulate factorization algorithms for which we can relate the errors introduced by the algorithms to perturbations, of an <u>a priori</u> specified magnitude, to the data of $\mathcal{P}$. The specific conditions under which the perturbations in the input data are acceptable depend on the problem and the aims of the computations, i.e., on the choice of $|\cdot|_n$. In fact, our theory gives more: a complete characterization of the structure of the perturbations to the original problem (cf. the proof of

Theorem 1 below). Indeed, the stability of the numerical algorithm will be related directly to the stability of the given problem $\mathcal{P}$ to perturbations of its data.

Definition 2. A factorization $\mathcal{F}(\pi,\Phi, ,Z,K)$ for problem $\mathcal{P}(B,F,U_\sigma,u_\alpha)$ is <u>bounded above</u> if there exist $M_1, M_2 > 0$ such that

$$(5) \qquad \|\Phi_\alpha^T(\cdot)\|_{(n,n_\alpha)} \leq M_\alpha, \ \alpha = 1,2.$$

$\mathcal{F}$ is <u>bounded below</u> if $\Phi_\alpha\Phi_\alpha^T$ is invertible for every $s \in [s_1,s_2]$ and there exist $m_1,m_2 > 0$ such that

$$(6) \qquad \|(\Phi_\alpha(s)\Phi_\alpha^T(s))^{-1}\|_{(n_\alpha,n_\alpha)} \leq \frac{1}{m_\alpha}, \ \alpha = 1,2,.$$

$\mathcal{F}$ is <u>bounded</u> if it is both bounded above and bounded below. ■

In order for a factorization to be useful for practical computations M_α and m_α should be numbers of "reasonable" magnitude. If M_α^2/m_α is large the condition number of $\Phi_\alpha\Phi_\alpha^T$ is large: Φ_α effectively loses rank. M_α and m_α are also related to the interpretation of the errors associated with the numerical realization of the factorization.

Indeed, suppose that $\sigma \in [\sigma_1,\sigma_2]$ is a target point in an interval over which we have the forward and backward factorizations $(\alpha = 1,2)$

$$(7a) \qquad \Phi_\alpha'(s) = -(\Phi_\alpha B)(s) + (Z_\alpha\Phi_\alpha)(s),$$

$$(7b) \qquad \Phi_\alpha(\sigma_\alpha) = U_\alpha;$$

and

(8a) $$\varphi_{\alpha}'(S) = -(\Phi_{\alpha}F)(s) + (Z_{\alpha}\,\varphi_{\alpha})(s),$$

(8b) $$\varphi_{\alpha}(s_{\alpha}) = u_{\alpha}.$$

Any numerical realizations Ψ_{α}, ψ_{α} of (7) and (8) may be viewed as the exact solutions of a perturbed problem

(9a) $$\Psi_{\alpha}'(s) = -(\Psi_{\alpha}B)(s) + (Z_{\alpha}\Psi_{\alpha})(s) + \Delta_{\alpha}(s),$$

(9b) $$\Psi_{\alpha}(\sigma_{\alpha}) = U_{\alpha} + V_{\alpha};$$

and

(10a) $$\psi'_{\alpha}(s) = -(\Psi_{\alpha}F)(s) + (Z_{\alpha}\psi_{\alpha})(s) + \delta_{\alpha}(s),$$

(10b) $$\psi_{\alpha}(\sigma_{\alpha}) = u_{\alpha} + v_{\alpha}.$$

In (9) and (10) the matrix Δ_{α} of size (n_{α},n) and vector δ_{α} of size $(n_{\alpha},1)$ represent the discretization errors of the numerical method used to solve (7) and (8). The matrices V_{α} and v_{α} represent the error in realizing the boundary conditions. We claim that $s \to v(s)$ defined on $[\sigma_1,\sigma_2]$ by

(11) $$\begin{bmatrix} \Psi_1(s) \\ \\ \Psi_2(s) \end{bmatrix} v(s) = \begin{bmatrix} \psi_1(s) \\ \\ \psi_2(s) \end{bmatrix}$$

is a δ-acceptable solution to

(12a) $$w'(s) = B(s)w(s) - F(s), \qquad s \in [\sigma_1,\sigma_2],$$

(12b) $$U_1 w(\sigma_1) = u_1, \quad U_1 w(\sigma_2) = u_2$$

with a δ determined solely by the conditioning of Ψ_α and the discretization errors.

Theorem 1. Suppose Ψ_α, $\alpha = 1,2$ solving (6) and (7) satisfy

(13a) $$|\Psi_\alpha(\cdot)|_{(n_\alpha,n)} \le M_\alpha,$$

(13b) $$|(\Psi_\alpha(\cdot)\Psi_\alpha^T(\cdot))^{-1}|_{(n_\alpha,n_\alpha)} \le \frac{1}{m_\alpha},$$

Then v defined by (11) is δ-acceptable (at $\sigma \in [\sigma_1,\sigma_2]$) with

(14a) $$\delta_B \le \max_{\alpha=1,2} \{\frac{M_\alpha}{m_\alpha} |\Delta_\alpha|_{(n_\alpha,n)}\},$$

(14b) $$\delta_F \le \max_{\alpha=1,2} \{\frac{M_\alpha}{m_\alpha} |\delta_\alpha|_{(n_\alpha,1)}\},$$

(14c) $$\delta_{U_\alpha} \le |V_\alpha|_{(n_\alpha,n)}$$

(14d) $$\delta_{u_\alpha} \le |v_\alpha|_{(n_\alpha,1)}.$$

In (14a) (14b) the norms as functions of s are assumed to be of L_p type.

Proof. Introduce matrix functions b_α by

(15) $$b_\alpha(s) = \Psi_\alpha^T(s)(\Psi_\alpha(s)\Psi_\alpha^T(s))^{-1}\Delta_\alpha(s);$$

and vector functions f_α by

$$f_\alpha(s) = \Psi^T_\alpha(s)(\Psi_\alpha(s)\Psi^T_\alpha(s))^{-1}\delta_\alpha(s). \tag{16}$$

It is easy to see that Ψ_α, ψ_α satisfy, for $\alpha = 1,2,$

$$\Psi'_\alpha(s) = -(\Psi_\alpha(B+b_\alpha))(s) + (Z_\alpha\Psi_\alpha)(s), \tag{17a}$$

$$\Psi_\alpha(\sigma_\alpha) = U_\alpha + V_\alpha, \tag{17b}$$

$$\psi'_{(s)} = -(\Psi_\alpha(F + f_\alpha))(s) + (Z_\alpha\Psi_\alpha)(s) \tag{18a}$$

$$\psi_\alpha(\sigma_\alpha) = u_\alpha, \tag{18b}$$

on $[\sigma_1,\sigma_2]$.

Now for fixed $\sigma \in [\sigma_1,\sigma_2]$ let

$$b(s;\sigma) = \begin{cases} b_1(s), & s \in [\sigma_1,\sigma], \\ b_2(s), & s \in [\sigma,\sigma_2], \end{cases}$$

and

$$f(s;\sigma) = \begin{cases} f_1(s), & s \in [\sigma_1,\sigma] \\ f_2(s), & s \in [\sigma,\sigma_2]. \end{cases}$$

Then $v = v(\sigma)$ satisfying (11) at $s = \sigma$ is the value at σ of the solution of the perturbed problem $P(B+b,\ F+f,\ U_\alpha+V_\alpha,\ u_\alpha+v_\alpha,\ \sigma_\alpha)$. The estimates (14) follow easily from (13), (15) and (16). ■

Remarks.

1. The perturbed problem P through which we interpret the approximate solution $v(\sigma)$ depends on the target point σ. The perturbation is different at every target point.

2. The magnitude of the perturbation $\delta = \{\delta_B, \delta_F, \delta_{U_\alpha}, \delta_{u_\alpha}\}$ is independent of the coefficients of the original problem, as long as a bound of the type (13) holds. In practice, it is the adaptive construction of the conditioning matrices Z_α which will guarantee (13).

3. We assume that the effect of roundoff errors is negligible. It could be incorporated into the matrices Δ_α and V_α, and into the vectors δ_α and v_α. The solution of the local systems (11) at each target point also introduces roundoff errors. We assume that the effects of these roundoff errors can be neglected with respect to the discretization errors which have already been made, especially if the computations are carried out in double precision. It is, however, possible to interpret also these roundoff errors as perturbations of the input data of the problem.

4. If we regard the errors in the realization of the boundary data as roundoff errors we may assume that $V_\alpha = 0$, $v_\alpha = 0$ in (9) and (10).

5. The expressions for b_α (15) and f_α (16) give the complete structure of the perturbations. This is in fact a stronger result than the claim of Theorem 1.

This, then, is what we mean by a stable factorization: a bounded factorization for which m_α and M_α are of such a priori known magnitudes that the approximate solution is δ-acceptable when the error tolerances of the numerical integrator for the factorization matrix/vector initial value problems are roughly on the order of δ.

2.5 Examples of stable factorizations

2.5.1 Continuous orthonormalization. Since in view of Theorem 2.4-1 we seek to control the behavior of $\Phi_\alpha\Phi_\alpha^T$, we study the properties of this product. Let $Q_\alpha(s) = \Phi_\alpha'(s)\Phi_\alpha^T(s)$ and observe that

(1) $$Q_\alpha(s) + Q_\alpha^T(s) = (\Phi_\alpha\Phi_\alpha^T)'(s).$$

Now since

(2) $$Q_\alpha(s) = -\Phi_\alpha B\Phi_\alpha^T + Z_\alpha\Phi_\alpha\Phi_\alpha^T$$

and Φ_α is of type (n_α,n), we can solve for Z_α:

(3) $$Z_\alpha = (\Phi_\alpha B\Phi_\alpha^T + Q_\alpha)(\Phi_\alpha\Phi_\alpha^T)^{-1}.$$

Then it is easy to prove

Lemma 1. Let $Z_\alpha(s)$ be determined by (3) for $s \to Q_\alpha(s)$ bounded, measurable, and anti-symmetric. Then $s \to \Phi_\alpha(s)$ satisfying

(4a) $$\Phi'_\alpha(s) = -\Phi_\alpha(s)B(s) + Z_\alpha(s)\Phi_\alpha(s)$$

(4b) $$\Phi_\alpha(s_\alpha) = U_\alpha$$

has the property that

(5) $$\Phi_\alpha(s)\Phi_\alpha^T(s) = U_\alpha U_\alpha^T.$$ ■

We can assume, without loss of generality, that $U_\alpha U_\alpha^T = Id_{n_\alpha}$, Indeed, we need only multiply U_α by $K_\alpha = (U_\alpha U_\alpha^T)^{-½}$. Thus we see that the factorization induced by the conditioning (3) with anti-symmetric Q_α

maintains the constraint that the rows of the transition matrix Φ_α are orthonormal.

However, in view of Theorem 2.4-1, we can expect to realize this constraint only approximately. Indeed, Z_α given by (3) for B will not do for the perturbed matrix $B + b_\alpha$. In practice, therefore, the approximate solution will drift away from the constraint manifold, and provision must be made for periodic discrete re-orthonormalizations. We refer to [19] for details and some numerical examples.

2.5.2 Stabilized continuous orthonormalization. This factorization is also based on (1), but with Q_α chosen to ensure asymptotic stability of the manifolds $\Phi_\alpha \Phi_\alpha^T = U_\alpha U_\alpha^T$.

Lemma 2. Let $Z_\alpha(s)$ be determined by (3) for Q_α given by

$$(6) \qquad Q_\alpha(s) = \kappa_\alpha\{U_\alpha U_\alpha^T - (\Phi_\alpha \Phi_\alpha^T(s)\}$$

where $\kappa_1 > 0$ and $\kappa_2 < 0$. Then the manifold $(\Phi_\alpha \Phi_\alpha^T)(s) = U_\alpha U_\alpha^T$ is asymptotically stable (forward in s for $\alpha = 1$ and backward in s for $\alpha = 2$).

Proof. It is enough to show that the solution $\Psi_\alpha(s) \equiv U_\alpha U_\alpha^T$ of

$$\Psi_\alpha'(s) = 2\kappa_\alpha(U_\alpha U_\alpha^T - \Psi_\alpha(s))$$

is asymptotically stable forward in s for $\alpha = 1$. But this is immediate since the coefficient matrix of this constant coefficient nonhomogeneous linear problem is $-2\kappa_\alpha \mathrm{Id}_{n_\alpha}$. ■

Remarks.

1. In an implementation of the continuous orthonormalization it might be tempting to replace $(\Phi_\alpha \Phi_\alpha^T)^{-1}$ in (3) by $(U_\alpha U_\alpha^T)^{-1}$ in order to avoid the expense of inverting $\Phi_\alpha \Phi_\alpha^T$ at every function evaluation. This of course invites disaster, as has been noted by Meyer [20] in a similar context. For a discussion of this method see also [10].

2. On the other hand, it is possible to show that for $|\kappa_\alpha|$ sufficiently large in (6), an implementation of stabilized continuous orthonormalization can afford to commit the above-mentioned crime. Unfortunately having $|\kappa_\alpha|$ large exacerbates any stiffness inherent in the problem.

2.5.3 Riccati factorization. The Ricatti factorization is based upon partitioning Φ_α in order to determine a maximally nonsingular $n_\alpha \times n_\alpha$ submatrix. We will need some notation and a few preliminary results.

For $0 < p < n$ we define matrices L_p and R_p of type (n,p) and $(n,n-p)$, respectively, by

$$L_p = \begin{bmatrix} Id_p \\ \hline 0 \end{bmatrix}, \qquad R_p = \begin{bmatrix} 0 \\ \hline Id_{n-p} \end{bmatrix}. \tag{7}$$

It is easy to verify

Lemma 3.

$$L_p L_p^T + R_p R_p^T = Id_n, \tag{8a}$$

$$L_p^T L_p = Id_p, \tag{8b}$$

(8c) $$R_p^T R_p = Id_{n-p},$$

(8d) $$[L_p \mid R_p] = Id_n.$$ ■

For a matrix A of size (n,m) and $0 < p < n$ we set

$$A_{1,} = L_p^T A, \qquad A_{2,} = R_p^T A$$

and for A of size (m,n) we set

$$A_1 = AL_p, \qquad A_2 = AR_p.$$

Evidently if A is of size (n,n) we will write A in partitioned block form as

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

where

$$A_{1,1} = L_p^T AL_p, \qquad A_{1,2} = L_p^T AR_p,$$

$$A_{2,1} = R_p^T AL_p, \qquad A_{2,2} = R_p^T AR_p.$$

Finally, if $\sigma = (\sigma_1,\ldots,\sigma_n)$ is a permutation of the integers $(1,\ldots,n)$ we denote by P^σ the corresponding $n \times n$ column permutation matrix; i.e., P^σ is the $n \times n$ identity matrix with columns permuted by σ. Then if U is of size (p,n) we have that UP^σ is a matrix the columns of which are those of U permuted by σ. If U is of size (n,m) then $(P^\sigma)^T U$ is a matrix the rows of which are those of U permuted by σ. Clearly, $(P^\sigma)^T P^\sigma = P^\sigma (P^\sigma)^T = Id_n$.

Lemma 4. If U is of type (p,n) with $0 < p \leq n$, then there exists a permutation σ such that the entries of the matrix $[(UP^{\sigma})L_p]^{-1}UP^{\sigma}$ are all in absolute value less than or equal to one.

Proof. Since U is of type (p,n) there exists a permutation τ such that $UP^{\tau}L_p$ is invertible. Let σ be such that

$$|\det(UP^{\sigma}L_p)| \geq |\det(UP^{\tau}L_p)|$$

for all permutations τ. Then by Cramer's rule the elements of $(UP^{\sigma}L_p)^{-1}UP^{\sigma}$ have the form

$$\frac{\det(UP^{\tau}L_p)}{\det(UP^{\sigma}L_p)}$$

for various choices of permutation τ. ■

We call $UP^{\sigma}L_p$ with σ as in Lemma 4 a maximally nonsingular (p,p) submatrix of U.

The following remarks apply both to the $\alpha = 1$ (left to right) and $\alpha = 2$ (right to left) factorizations. Accordingly, we suppress the subscript α and discuss only the left to right factorization.

We construct a solution to a problem of the following type: Given absolutely continuous functions $s \to B(s)$, $F(s)$ on $[s_1,s_2]$ with B of size (n,n) and F of size $(n,1)$, and given a matrix U of type (p,n) and vector u of size $(p,1)$, determine (piecewise absolutely continuous) functions $s \to \Phi(s)$, $\varphi(s)$ on $[s_1,s_2]$ such that Φ is of type (p,n), φ is of size $(p,1)$ and such that

$$(9) \qquad \Phi(s)w(s) = \varphi(s), \qquad s \in [s_1.s_2],$$

for any function $s \to w(s)$ satisfying

$$(10) \qquad \begin{cases} w'(s) = B(s)w(s) - F(s) \\ Uw(s_1) = u. \end{cases}$$

Step 0: Let $s^{(0)} = s_1$, $B^{(0)} = B$, $F^{(0)} = F$, $V^{(0)} = U$ and let $\sigma^{(1)}$ be such that $V^{(0)}P\sigma^{(1)}L_p$ is maximally non-singular. $\sigma^{(1)}$ can be obtained in practice by performing Gauss elimination with column pivoting on U. The result is the matrix

$$(11) \qquad \begin{aligned} U^{(1)} &= (V^{(0)}P\sigma^{(1)}L_p)^{-1} V^{(0)}P\sigma^{(1)} = [U_1^{(1)} \mid U_2^{(1)}] \\ &= [Id_p \mid U_2^{(1)}]. \end{aligned}$$

If the elimination is actually applied to the augmented matrix $[U \mid u]$, allowing only pivoting with respect to the first n columns, then we also obtain, with $v^{(0)} = u$,

$$(12) \qquad u^{(1)} = (V^{(0)}P\sigma^{(1)}L_p)^{-1} v^{(0)}.$$

Step i $(i > 0)$:

Suppose that $s^{(i-1)}$, $V^{(i-1)}$ of type (p,n), and $v(^{i-1)}$ of size $(p,1)$ are given. Let $\sigma^{(i)}$ be such that $(V^{(i-1)}P\sigma^{(i)}L_p)$ is maximally non-singular. Compute

$$(13) \qquad \begin{aligned} U^{(i)} &= (V^{(i-1)}P\sigma^{(i)}L_p)^{-1} V^{(i-1)}P\sigma^{(i)} \\ &= [U_1^{(i)} \mid U_2^{(i)}] = [Id_p \mid U_2^{(i)}] \end{aligned}$$

and

(14) $$u^{(i)} = (V^{(i-1)}P^{\sigma^{(i)}}L_p)^{-1}v^{(i-1)}.$$

Now set $B^{(i)} = (P^{\sigma^{(i)}})^T B^{(i-1)}P^{\sigma^{(i)}}$, $F^{(i)} = (P^{\sigma^{(i)}})^T F^{(i-1)}$ and let $\Phi^{(i)}$, $\varphi^{(i)}$ solve the initial value problems

(15a) $$\Phi^{(i)\prime}(s) = -(\Phi^{(i)}B^{(i)})(s) + (Z^{(i)}\Phi^{(i)})(s),$$

(15b) $$\Phi^{(i)}(s^{(i-1)}) = U^{(i)};$$

and

(16a) $$\varphi^{(i)\prime}(s) = -(\Phi^{(i)}F^{(i)})(s) + (Z^{(i)}\varphi^{(i)})(s),$$

(16b) $$\varphi^{(i)}(s^{(i-1)}) = u^{(i)},$$

where $Z^{(i)}$ is given by

(17) $$Z^{(i)}(s) = (\Phi_1^{(i)}B_{1,1}^{(i)} + \Phi_2^{(i)}B_{2,1}^{(i)})(s).$$

Recall that $\Phi_1^{(i)} = \Phi^{(i)}L_p$ and $\Phi_2^{(i)} = \Phi^{(i)}R_p$. It is not hard to see that (13) and (15) imply that

(18) $$\Phi_1^{(i)}(s) = Id_p$$

as long as (15a) holds. Moreover, $\Phi_2^{(i)}$ satisfies

(19a) $$\Phi_2^{(i)\prime}(s) = -B_{1,2}^{(i)} + B_{1,1}^{(i)}\Phi_2^{(i)} - \Phi_2^{(i)}B_{2,2}^{(i)} + \Phi_2^{(i)}B_{2,1}^{(i)}\Phi_2^{(i)},$$

(19b) $$\Phi_2^{(i)}(s^{(i-1)}) = U_2^{(i)},$$

while $\varphi^{(i)}$ satisfies

$$(20a) \qquad \varphi^{(i)\prime}(s) = -F_{1,}^{(i)} + B_{1,1}^{(i)}\varphi^{(i)} - \Phi_2^{(i)}F_{2,}^{(i)} + \Phi_2^{(i)}B_{2,1}^{(i)}\varphi^{(i)},$$

$$(20b) \qquad \varphi^{(i)}(s^{(i-1)}) = u^{(i)}.$$

Equations (19a) and (20a) constitute a (coupled) system of matrix Ricatti equations.

Of course, it may happen that the solution to (19), (20) does not exist on $[s^{(i-1)},s_2]$. It will, however, exist on some maximal interval $[s^{(i-1)},t)$, $t > s^{(i-1)}$. We do not expect to continue the numerical integration of (19), (20) all the way to t. Indeed, let $\Lambda > 1$ be an a priori given constant and set $d^{(i)}(s) = \|\Phi^{(i)}(s)\|_{(p,n)} = \|Id_p \quad \Phi_2^{(i)}(s)\|_{(p,n)}$. Let $t^{(i)} = \sup\{s\colon d^{(i)}(s) \le \Lambda d^{(i)}(s^{(i-1)})\}$ and then set $s^{(i)} = \min\{s_2,t^{(i)}\}$. We propagate the solution of (19), (20) only to $s = s^{(i)}$, at which point we set

$$V^{(i)} = \Phi^{(i)}(s^{(i)}), \qquad v^{(i)} = \varphi^{(i)}(s^{(i)}).$$

This completes Step i.

It is in this manner that we traverse the interval $[s_1,s_2]$ from left to right, adaptively constructing the (Riccati) factorization in such a way that it remains bounded in the sense of Definition 2.4-2. Indeed, we easily have that in the spectral norm $\|(\Phi(s)\Phi^T(s))^{-1}\| \le 1$ so that $\|(\Phi(s)\Phi^T(s))^{-1}\|_{(p,p)} \le \frac{1}{m(p)}$ while $\|\Phi^i(s)\|_{(\varphi,n)} \le M = \Lambda d$, where

$$d = \sup\{\|[Id_p \mid \Phi]\|_{(p,n)} : \Phi \text{ is of size } (p,n-p) \text{ and } |\Phi_{ij}| \le 1\}.$$

We refer to the points $s^{(i)}$, $i = 1,2,...$ as _switching_ or _resetting_ points.

Remarks.

1. The factorization (15), (16) with the conditioning matrix given in (17) again comprise a system of matrix ODE's on a manifold. However, the manifold is given _explicitly_ by (18). The advantages from a computational point of view are striking:

a. We are spared the integration of n_1^2 equations in the forward direction and n_2^2 equations in the backward direction.

b. The constraint manifold (18) is maintained explicitly.

2. Let us emphasize the very important point that the perturbation $\Delta_\alpha(s)$ of Section 2.4 is of the type $[0_p \mid \Delta]$. That is, we do not have perturbations in the first part of Δ_α because we explicitly maintain Φ_α in the form $[Id_p, \Phi]$. This point is not exploited by the general analysis which led to Theorem 4.1.

3. In connection with the Riccati equation approach we would like to mention the important, but generally unknown, papers by J. Taufer. See [22], [23].

We must, on the other hand, confront the difficulties associated with the numerical integration of matrix Riccati equations. We have already showed how to _exploit_ the possibility that the Riccati solution trajectories may "blow up" in finite time (i.e., by the use of our switching strategy). We turn in the next section to the development of a stable implicit solver for initial value problems for matrix Ricatti equations which exploits the special structure of the quadratic right hand side.

3 COMPUTATIONAL ASPECTS OF THE FACTORIZATION METHOD

Any numerical realization of the method of factorization should be robust, efficient, accurate, and stable. Moreover, it should exploit and preserve the special character of the particular factorization being used. These requirements pose special problems for the designer of factorization-based codes.

We have already noted the advantages of explicit over implicit constraint manifolds. There is also the question of the matrix character of the initial value problems of the factorization. It is tempting to use one of the excellent modern adaptive initial value codes now available. This, however, means that the matrices must be "unrolled" into equivalent vector form. This poses no special problems except in cases --usually the ones of practical interest--in which a stiff (implicit) solver is required. Propagation of size (n_α,n) transition matrix in unrolled form with a stiff solver requires the computation and repeated decomposition of Jacobian matrices of size $(n_\alpha n,n_\alpha n)$. This represent a severe computational burden even for moderate n_α and n ($n_\alpha = 10$ and $n = 20$, say).

Indeed, this computational burden is so severe that it is part of the folklore of control engineering--where matrix Riccati equations play a central role--that one should not integrate the matrix Riccati initial value problem in order to determine steady-state solutions. Special methods have been developed to solve the algebraic (steady-state) Riccati equation instead ([14], [21]).

We will show in this section how these computational limitations can be overcome by the design of special matrix initial value solvers which exploit the structure of the factorization equations. We analyze, in particular, the accuracy and numerical stability of a solver for matrix initial value problems of Riccati type. We will need the concept of a numerical process (cf. [4]).

3.1 Numerical processes

Denote by $M_{(n,m)}$ the vector space of matrices of size (n,m) endowed with a norm $|\cdot|$, and consider the matrix initial value problem

(1a) $$\Phi'(s) = R(\Phi(s),s),$$

(1b) $$\Phi(s_0) = \Phi_0,$$

with $[a,b] \ni s \to \Phi(s) \in M_{(n,m)}$ and $R: M_{(n,m)} \times [a,b] \to M_{(n,m)}$ a Lipschitz continuous mapping.

Definition 1. A <u>one step numerical process</u> in $M_{(n,m)}$ is a sequence $\{\Phi_i\} \subset M_{(n,m)}$ and a sequence $\{h_i\} \subset \mathbf{R}$ such that

(2) $$\Phi_{i+1} = P_i(\Phi_i,h_i),$$

where Φ_0 and the sequence $\{h_i\}$ are given, and where each P_i is a mapping

$$P_i: M_{(n,m)} \times \mathbf{R} \to M_{(n,m)}. \qquad \blacksquare$$

We regard the process (2) as a discrete approximation to the continuous process(1) if it is consistent; that is,

Definition 2. The numerical process (2) is <u>consistent</u> to order p with (1) if

(3a) $$P_i(\Phi_i,h_i) = \hat{\Phi}(s_i + h_i) + O(h_i^{p+1})$$

for

(3b) $$s_i = s_0 + \sum_{j=0}^{i-1} h_j,$$

and $\hat{\Phi}(s)$ satisfies (1a) with initial condition $\hat{\Phi}(s_i) = \Phi_i$. If in addition the form (2) is regular, i.e.,

$$|P_i(\Phi_i,h_i) - P_i(\bar{\Phi}_i,h_i)| \leq M|\Phi_i-\bar{\Phi}_i|$$

for $\bar{\Phi}_i \in M_{n,m}$ with $|\Phi_i - \bar{\Phi}_i| < \varepsilon$ for all i and with M independent of i, then (2) is a general one step method for the solution of (1a,b) which converges with the rate $O(h^p)$ (cf. [4], Section 3.3.1). ■

Another way to say this is that (1) is the (order p) closure [5] of the numerical process (2).

Definition 3. The one step numerical process (2) is bounded if there exists $M \geq 0$ such that

(4) $$|\Phi_i| \leq M$$

and finally,

Definition 4. The one step numerical process (2) is stable if there exists $L \leq 1$ and $\varepsilon_0 > 0$ such that

(5) $$|\Phi_{i+1} - \tilde{\Phi}_{i+1}| \leq L|\Phi_i - \tilde{\Phi}_i|$$

whenever

(6a) $$|\Phi_0 - \tilde{\Phi}_0| \leq \varepsilon_0$$

and

(6b) $$\tilde{\Phi}_{i+1} = P_i(\tilde{\Phi}_i, h_i).$$

If $L < 1$ then (2) is *strongly stable*.

Definition 4 is an adaptation of the notion of BN-stability [9]. In contrast to the concepts of A-stability and B-stability, the definition is made without reference to a particular test equation.

3.2 A one-step process for Ricatti equations

Consider the special case of (1) given by (cf. equations (2.5-29))

(1a) $$\Psi'(s) = -D(s) + A(s)\Psi(s) - \Psi(s)B(s) + \Psi(s)C(s)\Psi(s)$$

(1b) $$\Psi(s_0) = \Psi_0.$$

Here $\Psi(s) \in M_{(n,m)}$, $A(s) \in M_{(n,n)}$, $B(s) \in M_{(m,m)}$, $C(s) \in M_{(m,n)}$, and $D(s) \in M_{(n,m)}$. Let us use the notation $A_{i+\alpha} = A(s_i + \alpha h_i)$, for $\alpha \in [0,1]$. Then we define a one step numerical process for (1) in two stages:

(2a) $$\Psi_{i+\frac{1}{2}} = [Id_n - \tfrac{1}{2} h_i(A_{i+\frac{1}{2}} + \Psi_i C_{i+\frac{1}{2}})]^{-1} \cdot [\Psi_i - \tfrac{1}{2} h_i(\Psi_i B_{i+\frac{1}{2}} + D_{i+\frac{1}{2}})]$$

(2b) $$\Psi_{i+1} = [\Psi_{i+\frac{1}{2}} + \tfrac{1}{2} h_i(A_{i+\frac{1}{2}} \Psi_{i+\frac{1}{2}} - D_{i+\frac{1}{2}})] \cdot [Id_m + \tfrac{1}{2} h_i(B_{i+\frac{1}{2}} - C_{i+\frac{1}{2}} \Psi_{i+\frac{1}{2}})]^{-1}.$$

Remarks.

1. It is certainly possible to eliminate the intermediate solution $\Psi_{i+\frac{1}{2}}$ between (2a) and (2b) in order to write (2) in the form (3.1-2).

2. We evaluate the matrix coefficients A, B, C, D only once, at the midpoint, $s_{i+\frac{1}{2}}$, of the step. The RHS of (1) is not evaluated at all.

3. There are two matrix decompositions per step; one of type (n,n) and the other of type (m,m), for an operation count of $O(n^3 + m^3)$. This is very favorable compared to the $O(nm)^3$ operations required for the factorization of the Jacobian of the RHS of (1a) when written in unrolled (vector) form.

Of course, in an implementation of the Riccati factorization in order to solve BVP's we must propagate a transition vector, ψ, as well as the transition matrix, Ψ. In addition to (1), we have (cf. equations 2.5-20)

$$\text{(3a)} \qquad \psi'(s) = -F(s) + A(s)\psi(s) - \Psi(s)G(s) + \Psi(s)C(s)\psi(s),$$

$$\text{(3b)} \qquad \psi(s_0) = \psi_0,$$

in which $\psi(s) \in M_{(n,1)}$, $F(s) \in M_{(n,1)}$, $G(s) \in M_{(m,1)}$, and A and C are as above. The integration of (3) is done simultaneously with that of (1) by observing that (1), (3) can be written

$$\begin{aligned} [\psi \mid \psi]' &= -[D \mid F] + A[\Psi \mid \psi] \\ \text{(4)} \qquad & -[\Psi \mid \psi] \left[\begin{array}{c|c} B & G \\ \hline 0 & 0 \end{array}\right] + [\Psi \mid \psi]\left[\begin{array}{c} C \\ \hline 0 \end{array}\right][\Psi \mid \psi] \end{aligned}$$

which is again of the type (1). We apply the scheme (2) to the augmented system (4) in practice. Of course, the zero coefficients are not stored in the computer implementation of (2) applied to (4).

Remark.

If $F \in M_{(n,p)}$ and $G \in M_{(m,p)}$, then $\psi \in M_{(n,p)}$ in equations (3). This corresponds to solving the original BVP (2.1-1) for p right hand sides. This can be accomodated trivially in (2) (4).

3.3 Consistency of the method

The numerical process (3.2-2) is consistent to order 2 for (3.2-1). Indeed (3.2-1) is of the type

$$(1) \qquad \begin{cases} \Psi' = f(s;\Psi,\Psi), \\ \Psi(s_0) = \Psi_0, \end{cases}$$

while the corresponding discrete process (3.2-2) is of the type

$$(2a) \qquad \Psi_{i+\frac{1}{2}} = \Psi_i + \frac{h_i}{2} f(s_{i+\frac{1}{2}};\Psi_i,\Psi_{i+\frac{1}{2}})$$

$$(2b) \qquad \Psi_{i+1} = \Psi_{i+\frac{1}{2}} + \frac{h_i}{2} f(s_{i+\frac{1}{2}};\Psi_{i+1},\Psi_{i+\frac{1}{2}}).$$

In order to analyze the one-step error of (2), it is sufficient to consider the case $i = 0$; and for notational clarity we write $h = h_0$ and suppress the dependence upon $s_{\frac{1}{2}}$ in (1) and (2). The exact solution satisfies

$$(3) \qquad \Psi(s_0+h) = \Psi_0 + h\Psi_0' + \frac{h^2}{2}\Psi_0'' + O(h^3)$$

where

(4a) $$\Psi_0' = f(\Psi_0,\Psi_0)$$

(4b) $$\Psi_0'' = g_1(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0)$$

(4c) $$g_1(\Psi_0,\Psi_0) = \frac{\partial}{\partial\Phi} f(\Phi,\Psi)\Big|_{\Phi=\Psi_0,\Psi=\Psi_0}$$

(4d) $$g_2(\Psi_0,\Psi_0) = \frac{\partial}{\partial\Psi} f(\Phi,\Psi)\Big|_{\Phi=\Psi_0,\Psi=\Psi_0}.$$

On the other hand, from (2) we have that

(5) $$\begin{aligned} \Psi_1 &= \Psi_0 + \frac{h}{2}[f(\Psi_0,\Psi_{\frac{1}{2}}) + f(\Psi_1,\Psi_{\frac{1}{2}})] \\ &= \Psi_0 + \frac{h}{2}[f(\psi_0,\psi_0) + \frac{h}{2} g_2(\Psi_0,\Psi_0)f(\psi_0,\Psi_{\frac{1}{2}}) \\ &\quad + f(\Psi_{\frac{1}{2}},\Psi_{\frac{1}{2}}) + \frac{h}{2} g_1(\Psi_{\frac{1}{2}},\Psi_{\frac{1}{2}})f(\Psi_1,\Psi_{\frac{1}{2}})] \\ &\quad + O(h^3). \end{aligned}$$

Thus

(6) $$\begin{aligned} \Psi(s_0+h) - \Psi_1 &= hf(\Psi_0,\Psi_0) - \frac{h}{2} f(\Psi_0,\Psi_0) - \frac{h}{2} f(\Psi_{\frac{1}{2}},\Psi_{\frac{1}{2}}) \\ &\quad + \frac{h^2}{2}[g_1(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0)] \\ &\quad - \frac{h^2}{4}[g_1(\Psi_{\frac{1}{2}},\Psi_{\frac{1}{2}})f(\Psi_1,\Psi_{\frac{1}{2}}) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_{\frac{1}{2}})] \\ &\quad + O(h^3). \end{aligned}$$

However,

$$\begin{aligned} f(\Psi_{\frac{1}{2}},\Psi_{\frac{1}{2}}) &= f(\Psi_0,\Psi_0) + \frac{h}{2}[g_1(\Psi_0,\Psi_0)f(\Psi_0,\Psi_{\frac{1}{2}}) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_{\frac{1}{2}})] \\ &\quad + O(h^2). \end{aligned}$$

Therefore, we have

$$(7)\qquad \Psi(s_0+h) - \Psi_1 = -\frac{h^2}{4}[g_1(\Psi_0,\Psi_0)f(\psi_0,\Psi_{1/2}) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_{1/2})]$$

$$+\frac{h^2}{2}[g_1(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0) + g_2(\Psi_0,\Psi_0)f(\Psi_0,\Psi_0)]$$

$$-\frac{h^2}{4}[g_1(\Psi_{1/2},\Psi_{1/2})f(\Psi_1,\Psi_{1/2}) + g_2(\Psi_0,\Psi_{1/2})]$$

$$+ O(h^3).$$

But also, we note that

$$f(\Psi_0,\Psi_{1/2}) = f(\Psi_0,\Psi_0) + O(h),$$

$$g_1(\Psi_{1/2},\Psi_{1/2}) = g_1(\Psi_0,\Psi_0\} + O(h),$$

$$f(\Psi_1,\Psi_{1/2}) = f(\Psi_0,\Psi_0) + O(h),$$

so that up to terms of order $O(h)$, the bracketed quantities in (7) sum to zero:

$$(8)\qquad \Psi(s_0+h) - \Psi_1 = O(h^3).$$

In practice, the 2nd order process (3.2-2) can be extrapolated to higher order, with the order and the step size h varied adaptively based upon comparisons of consecutive stages in the extrapolation. An implementation at the University of Maryland uses a fixed number of extrapolation stages, three; it is a method of type 3-4 with adaptive step selection.

The step control is based on the usual simultaneous usage of methods of two different orders. The local error is controlled so that the method uses the minimal number of steps when the perturbation is measured in the norm L_∞ or on the norm L_1 (see [7]). In [7] it has also been shown that the error-per-unit step approach is optimal with respect to the L_∞ norm and the error per-step is optimal with respect to the L_1 norm. We give some numerical examples in Section 4. But first let us analyze the stability of the numerical processes (3.2-2,4).

3.4 Stability of the method

The boundedness and stability of the process (3.2-2) are inherited from the continuous processes. If the solution to (3.2-1) exists on $[s_1,s_2] \ni s_0$, then it is uniformly bounded on $[s_1,s_2]$. Suppose further that the solution is locally stable forward in s. Then it is not hard to see that

Lemma 1. If $s \to \Psi(s)$ satisfies (1) on $[s_1,s_2] \ni s_0$, then the trajectory $\Psi(s)$ is locally stable forward in s if and only if

$$\text{(1)} \qquad \operatorname{Re} \lambda(s) - \operatorname{Re} \mu(s) \le 0$$

where $\lambda(s)$ is any eigenvalue of $F(s) \equiv A(s) + \Psi(s)C(s)$ and $\mu(s)$ is any eigenvalue of $G(s) \equiv B(s) - C(s)\Psi(s)$.

Proof. We need only observe that (1) is necessary and sufficient for the trivial solution of the linearized problem,

$$\text{(2)} \qquad \Phi'(s) = [A(s) + \Psi(s)C(s)]\Phi(s) - \Phi(s)[B(s) - C(s)\Psi(s)],$$

to be locally stable forward in s.

Indeed, writing (2) in unrolled vector form we analyze the eigenvalues of the Kronecker sum of the matrices $F(s)$ and $G(s)$ (see [8], p. 230). The eigenvalues of this sum take exactly the form (1). ∎

Corollary 2. The trajectory Ψ is locally stable backward in s if and only if

$$\text{Re } \lambda(s) - \text{Re } \mu(s) \geq 0. \tag{3}$$ ∎

This stability property is inherited by the discrete process (3.2-2). Indeed, let

$$\Phi_i = \Psi_i - \tilde{\Psi}_i \tag{4}$$

where $\tilde{\Psi}_i$ solves (3.2-2) with initial value $\tilde{\Psi}_0$. Then we have

$$\Phi_{i+\frac{1}{2}} = \Phi_i + \frac{h_i}{2}[A_{i+\frac{1}{2}} \Phi_{i+\frac{1}{2}} - \Phi_i B_{i+\frac{1}{2}} + \Phi_i C_{i+\frac{1}{2}} \tilde{\Psi}_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}} C_{i+\frac{1}{2}} \Phi_{i+\frac{1}{2}}] \tag{5a}$$

$$\Phi_{i+1} = \Phi_{i+\frac{1}{2}} + \frac{h_i}{2}[A_{i+\frac{1}{2}}\Phi_{i+\frac{1}{2}} - \Phi_{i+1}B_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}} C_{i+\frac{1}{2}} \Phi_{i+\frac{1}{2}} + \Phi_{i+1}C_{i+\frac{1}{2}} \tilde{\Psi}_{i+\frac{1}{2}}]. \tag{5b}$$

Upon eliminating $\Phi_{i+\frac{1}{2}}$ we have

(6)

$$\Phi_{i+1} = [Id_n + \frac{h_i}{2}(A_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}} C_{i+\frac{1}{2}})][Id_n - \frac{h_i}{2}(A_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}} C_{i+\frac{1}{2}})]^{-1} \cdot$$
$$\cdot \Phi_i [Id_m - \frac{h_i}{2}(B_{i+\frac{1}{2}} - C_{i+\frac{1}{2}} \tilde{\Psi}_{i+\frac{1}{2}})][Id_m + \frac{h_i}{2}(B_{i+\frac{1}{2}} - C_{i+\frac{1}{2}} \tilde{\Psi}_{i+\frac{1}{2}})]^{-1}.$$

Now set

(7) $$F_i = A_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}} C_{i+\frac{1}{2}} , \qquad G_i = B_{i+\frac{1}{2}} - C_{i+\frac{1}{2}} \tilde{\Psi}_{i+\frac{1}{2}}$$

and rewrite (6) as

(8) $$\Phi_{i+1} = (Id_n + \frac{h_i}{2} F_i)(Id_n - \frac{h_i}{2} F_i)^{-1} \Phi_i (Id_m - \frac{h_i}{2} G_i)(Id_m + \frac{h_i}{2} G_i)^{-1} .$$

We therefore have that

(9) $$|\Phi_{i+1}| \leq \Gamma_i |\Phi_i| ,$$

where

(10a) $$\Gamma_i = \max|\gamma_i| ,$$

and

(10b) $$\gamma_i = \frac{(1 + \frac{h_i}{2} \lambda_i)}{(1 - \frac{h_i}{2} \lambda_i,} \, \frac{(1 - \frac{h_i}{2} \mu_i)}{(1 + \frac{h_i}{2} \mu_i)} .$$

In (10), the maximum is taken over γ_i of the form (106) where λ_i is any eigenvalue of F_i and μ_i is any eigenvalue of G_i.

For numerical stablity of (3.2-2) we must ensure that $|\gamma_i| \leq 1$ ($|\gamma_i| < 1$ for strong stability). Let us compute, then, $|\gamma_i|^2$. We drop the index i for notational clarity.

(11) $$|\gamma|^2 = \frac{(1 + \frac{h}{2} \lambda)(1 - \frac{h}{2} \bar{\lambda}) \; (1 - \frac{h}{2} \mu)(1 - \frac{h}{2} \bar{\mu})}{(1 - \frac{h}{2} \lambda)(1 - \frac{h}{2} \bar{\gamma}) \; (1 + \frac{h}{2} \mu)(1 + \frac{h}{2} \bar{\mu})} = \frac{1 + \frac{h}{2} p(\frac{h}{2})}{1 + \frac{h}{2} q(\frac{h}{2})} .$$

Here

$$(12) \qquad p(h) = 2\,\mathrm{Re}(\lambda-\mu) + h(|\lambda|^2 + |\mu|^2 - 4\,\mathrm{Re}\,\lambda\,\mathrm{Re}\,\mu)$$
$$+ 2h^2(|\mu|^2\mathrm{Re}\,\lambda - |\lambda|^2\,\mathrm{Re}\,\mu) + h^3|\lambda|^2|\mu|^2$$

and

$$(13) \qquad q(h) = 2\,\mathrm{Re}(\mu-\lambda) + h(|\lambda|^2 + |\mu|^2 - 4\,\mathrm{Re}\,\lambda\,\mathrm{Re}\,\mu)$$
$$+ 2\,h^2(|\lambda|^2\,\mathrm{Re}\,\mu - |\mu|^2\,\mathrm{Re}\,\lambda) + h^3|\lambda|^2|\mu|^2.$$

Stability is guaranteed whenever

$$(14) \qquad \frac{h}{2}\,r(\frac{h}{2}) = \frac{h}{2}\,[p(\frac{h}{2}) - q(\frac{h}{2})] \leq 0$$

with strict inequality for strong stability. Of course, the representation

$$(15) \qquad r(\frac{h}{2}) = 4\,\mathrm{Re}(\lambda-\mu) + h^2(|\mu|^2\,\mathrm{Re}\,\lambda - |\lambda|^2\,\mathrm{Re}\,\mu)$$

gives us

Lemma 3. If $\{\Psi_i\}$ and $\{\tilde{\Psi}_i\}$ are discrete trajectories satisfying (3.2-2) and such that $\mathrm{Re}\,\lambda_i \leq 0$, and $\mathrm{Re}\,\mu_i \geq 0$ where λ_i is an eigenvalue of $A_{i+\frac{1}{2}} + \Psi_{i+\frac{1}{2}}C_{i+\frac{1}{2}}$ and μ_i is an eigenvalue of $B_{i+\frac{1}{2}} - C_{i+\frac{1}{2}}\tilde{\Psi}_{i+\frac{1}{2}}$, then the process (3.2-2) is stable forward $(h_i > 0)$. If $\mathrm{Re}\,\lambda_i \geq 0$ and $\mathrm{Re}\,\mu_i \leq 0$ then the process (3.2-2) is stable backward $(h_i < 0)$.

Lemma 4. Let λ_i and μ_i be as in Lemma 1, but with the weaker hypothesis $\mathrm{Re}(\lambda_i-\mu_i) \leq 0$. Then the process (3.2-2) is stable forward $(h_i > 0)$ for

(16)

$$h_i \le h_{max} = \begin{cases} \min\left[\dfrac{4\ \mathrm{Re}(\mu_i-\lambda_i)}{|\mu_i|^2\mathrm{Re}\ \lambda_i - \tfrac{1}{2}\ \lambda_i|^2\mathrm{Re}\ \mu_i}\right]^{\frac{1}{2}}, & |\mu_i|^2\mathrm{Re}\ \lambda_i - |\lambda_i|^2\mathrm{Re}\ \mu_i > 0 \\ \\ \infty & \text{otherwise.} \end{cases}$$

The process is stable backward $(h_i < 0)$ for $\mathrm{Re}(\lambda_i-\mu_i) \ge 0$ and

(17)

$$-h_i \le h_{max} = \begin{cases} \min\left[\dfrac{4\ \mathrm{Re}(\lambda_i-\mu_i)}{|\lambda_i|^2\mathrm{Re}\ \mu_i - |\mu_i|^2\mathrm{Re}\ \lambda_i}\right]^{\frac{1}{2}}, & |\mu_i|^2\mathrm{Re}\ \lambda_i - |\lambda_i|^2\mathrm{Re}\ \mu_i < 0 \\ \\ \infty & \text{otherwise.} \end{cases}$$ ■

Remarks.

1. Lemma 3 is the matrix ODE analog of A-stability in the constant coefficient linear case $(C(s) = 0)$.

2. If $|\mu|^2\mathrm{Re}\ \lambda - |\lambda|^2\mathrm{Re}\ \mu > 0$ while $\mathrm{Re}\ \lambda - \mathrm{Re}\ \mu < 0$ in Lemma 4, it is interesting to examine how h_{max} depends on λ. For example, suppose $\lambda = \alpha\mu$ with $0 < \alpha < 1$. Then $h_{max} = \dfrac{2}{|\mu|}\alpha^{-\frac{1}{2}}$. Thus as long as the stronger hypotheses of Lemma 1 are not violated too severely (α near 1) the stability limit h_{max} on the step-size will still be quite large.

3. For a well-posed elliptic boundary value problem it can be shown that $\mathrm{Re}(\text{e.v. } F) \le 0$ and $\mathrm{Re}(\text{e.v. } G) \ge 0$ for the forward process and $\mathrm{Re}(\text{e.v. } F) \ge 0$ and $\mathrm{Re}(\text{e.v. } G) \le 0$ for the backward process . . . exactly the conditions of Lemma 1. This will also be the case for the Riccati equations which arise in classical linear-quadratic optimal control. These are usually posed backward and have the additional property that $-G^T = F > 0$.

4 NUMERICAL EXAMPLES

In this section we give several examples of the performance of a factorization based two point boundary value code based on the ideas presented above. We have chosen some examples to illustrate the robustness and effectiveness of the solver on problems stemming from engineering. We include also an unstable turning point problem for which our method fails.

We have already noted how the independent and parallel structure of the forward and backward factorizations coupled with the solution of local linear systems at the target points combine to minimize storage and the computational burden of the method. The adaptive mesh selection is based on a single solution of the problem, in contrast to the multiple-pass approach to mesh refinement used in global methods such as finite differences, collocation, or finite elements. This feature also greatly reduces the computational costs.

The numerical examples illustrate these features, but highlight the performance of the method on problems having a singular perturbation character, i.e., problems the solutions of which exhibit boundary or interior layers. We show that such problems can be solved effectively <u>without special handling</u> such as upwinding or asymptotic expansions.

4.1 A stable singular perturbation problem

Consider the problem

(1a) $$\varepsilon u''(s) + u'(s) = 1$$

(1b) $$u(0) = u(1) = 0$$

the solution of which is

$$(2) \qquad u(s) = s - \frac{1-\exp(\frac{s}{\varepsilon})}{1-\exp(-\frac{1}{e})} .$$

There is an $O(\varepsilon)$ boundary layer at $s = 0$.

There are a number of ways that (1) can be cast into the first order form 2.1-1; we explore some of the theoretical and computational implications of such re-formulations.

Method 1 (M1)

Let $w^T = (w_1, w_2)$ be defined by

$$(3) \qquad w_1 = u, \qquad w_2 = u'$$

and obtain

$$(4a,b) \qquad B = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{1}{\varepsilon} \end{bmatrix}, \qquad F = \begin{bmatrix} 0 \\ -\frac{1}{\varepsilon} \end{bmatrix} .$$

$$(4c,d) \qquad U_1 = U_2 = [1 \quad 0], \qquad u_1 = u_2 = 0.$$

Method 2 (M2)

Let $w^T = (w_1, w_2)$ be defined by

$$(5) \qquad w_1 = u, \qquad w_2 = \varepsilon u'$$

and obtain

$$(6a,b) \qquad B = \begin{bmatrix} 0 & \frac{1}{\varepsilon} \\ 0 & -\frac{1}{\varepsilon} \end{bmatrix}, \qquad F = \begin{bmatrix} 0 \\ -1 \end{bmatrix} .$$

$$(6c,d) \qquad U_1 = U_2 = [1 \quad 0], \qquad u_1 = u_2 = 0.$$

(M1) and (M2) have the form

$$w' = Bw - F, \tag{7a}$$

$$U_1w(0) = U_2w(1) = 0. \tag{7b}$$

The two possibilities for the Riccati factorization (which are adaptively alternated) corresponding to (7) are factorization F1:

$$\Phi_1 = \mathrm{Id}, \tag{8a}$$

$$\Phi_2' = -B_{12} + B_{11}\Phi_2 - \Phi_2B_{22} + \Phi_2B_{21}\Phi_2, \tag{8b}$$

$$\varphi' = -F_1 + B_{11} \quad - \Phi_2F_2 + \Phi_2B_{21} \ , \tag{8c}$$

and factorization F2:

$$\Phi_1' = -B_{21} + B_{22}\Phi_1 - \Phi_1B_{11} + \Phi_1B_{12}\Phi_1, \tag{9a}$$

$$\Phi_2 = \mathrm{Id}, \tag{9b}$$

$$\varphi' = -F_2 + B_{22} \quad - \Phi_1F_1 + \Phi_1B_{12} \ , \tag{9c}$$

where $\Phi = [\Phi_1 \ \Phi_2]$ and φ are the transition matrix and transition vector, respectively.

Suppose that matrix B and vector F are perturbed by b and f, respectively. Then the solution w to (7) is perturbed by v satisfying

$$v' = Bv + (bw - f) \tag{10a}$$

$$U_1v(0) = U_2v(1) = 0. \tag{10b}$$

The matrix b and the vector f arise computationally due to the discretization errors involved in solving (8) or (9). In the case of F1, if the discretization error of (8b) is $-b_{12}$ and that of (8c) is $-f_1$, then we see that

$$b = \begin{bmatrix} 0 & b_{12} \\ 0 & 0 \end{bmatrix}, \tag{11a}$$

$$f = \begin{bmatrix} f_1 \\ 0 \end{bmatrix}. \tag{11b}$$

For the factorization F2 the corresponding perturbations are

$$b = \begin{bmatrix} 0 & 0 \\ b_{21} & 0 \end{bmatrix}, \tag{12a}$$

$$f = \begin{bmatrix} 0 \\ f_2 \end{bmatrix} \tag{12b}$$

where now $-b_{21}$ is the discretization error of (9a) and $-f_2$ is the discretization error of (9c). (Eqns (8a) and (9b) are "solved " exactly.)

Since for M1, $w_2 = O(1/\varepsilon)$ in the boundary layer, we see that $b_{12} = O(\Delta)$ perturbations in B_{12} can lead to large $O(\Delta/\varepsilon)$ errors in v. We therefore would like to ensure that $b_{12} = O(\Delta\varepsilon)$ in the layer in order to get $v = O(\Delta)$. This can be done by brute force using an integration tolerance $\tau = \Delta\varepsilon$. This strategy will cause the adaptive solver to use more steps globally as well as in the boundary layer. The problem does not arise in factorization F2 since the perturbation has been shifted to b_{21}, the coefficient of $w_1 = O(1)$ in (10a). Moreover, although the factorization F1 is in effect at the start of both forward and backward sweeps due to the structure of the boundary condition matrices U_1 and U_2, the rapid growth of Φ_2 (for $\varepsilon << 1$) causes a switch to F2 after only a few steps.

Now consider the effect of measuring the error by the norms

(13a) $$\|v\|_1 = \sup_s\{|v_1(s)| + |v_2(s)|\},$$

(13b) $$\|v\|_2 = \sup_s\{|v_1(s)| + \varepsilon|v_2(s)|\}.$$

These vector norms induce corresponding matrix norms on b given by

$$\|b\|_1 = \sup_s \max[|b_{11}| + |b_{21}|, [b_{12}| + |b_{22}|]$$

(14b) $$\|b\|_2 = \sup_s \max[|b_{11}| + \varepsilon|b_{21}|, \varepsilon^{-1}|b_{12}| + |b_{22}|].$$

For factorization F1, keeping $\|b\|_2 < \Delta$ means exactly that $|b_{12}| < \Delta\varepsilon$. Consider a transformed problem

(15) $$\tilde{w}' = \tilde{B}\tilde{w} - \tilde{F}$$

for which $\|\tilde{b}\|_1 = \|b\|_2$: let $A = \begin{bmatrix}1 & 0\\ 0 & \varepsilon\end{bmatrix}$, and set $\tilde{w} = Aw$, $\tilde{B} = ABA^{-1}$,and $\tilde{F} = AF$. Then it is easy to see that if (7) is solved with $\tilde{B}$, $\tilde{F}$ replacing B and F, then $\tilde{b}_{12} = \varepsilon^{-1}b_{12}$. Since for F1, $\|\tilde{b}\|_1 = \sup_s|\tilde{b}_{12}| = \sup_s|\varepsilon^{-1}b_{12}| = \|b\|_2$ we see that solving F1 with tolerances Δ, δ for (15) is equivalent to solving (7) with tolerances $\varepsilon\Delta$, $\varepsilon\delta$. Of course

$\tilde{B}$ and $\tilde{F}$ lead exactly to problem M2.

We have shown, then, that transformations of the original problem (1) into various 1st order forms can be interpreted as selecting a norm for the original problem, and affect the interpretation of the perturbations of the coefficients caused by discretization errors in the solution of the factorization initial value problems.

We further consider four other first order problems computing $w = (w_1, w_2)^T$:

Method 3 (M3)

(16) $$w_1 = \varepsilon u' + u, \qquad w_2 = u';$$

Method 4 (M4)

(17) $$w_1 = u, \qquad w_2 = \varepsilon u' - u;$$

Method 5 (M5)

(18) $$w_1 = u, \qquad w_2 = \varepsilon u' + u;$$

Method 6 (M6)

(19) $$w_1 = u, \qquad w_2 = \varepsilon u' + (s - \tfrac{1}{2}).$$

We are interested in the computation of $u'(0)$ by the mentioned six formulations. Therefore only one target point is considered, namely $s = 0$. Hence, only integration from right to left has to be performed by the factorization method. The computation has been in double precision and "per step" step selection criterion. The initial step was taken to be $h = \varepsilon$. Table 4.1-4.6 show the error in $u'(0)$ obtained by the method (M1)-(M-6) for various ε and tolerances τ. In addition, the number of integrations steps is given in the tables.

Comparing Tables 4.1 and 4.2 we see that the methods (M1) and (M2) produce identical results if $\tau_1 = \varepsilon\tau_2$ where by τ_i the tolerances used in method M_i were denoted. This is directly related to the analysis mentioned above. Methods (M1) and (M3) are using as one of the variables $w_2 = u'$, while (M2), (M4), (M5), (M6) are using the variable w_2 involving $\varepsilon u'$. This leads to similar performances of these two groups of methods. Nevertheless, there are some differences (see (M2) and (M6)) which are caused by the different structure of the perturbations in B and F. Results shown in Table 4.1 are heavily influenced by low tolerances τ for which the number of steps is independent of τ. This disappears for τ smaller as can be seen in Table 4.2, which is, as we said, essentially the method (M1) with tolerance $\tau\varepsilon$. We see in Table 4.2 that the error in $u'(0)$ for fixed ε is proportional to τ. This is because the perturbations in the input data caused by the approximate solution of the solved ODEs are of magnitude τ.

From (10) we thus expect that the error in $u'(0)$ is of order $\frac{\tau}{\varepsilon}$. We see this character from Table 4.2 especially for small τ and ε (when we are in the asymptotic range). For a particular value of τ ($\tau = 10^{-8}$) there is a (local) increase in accuracy. This is the effect of some cancelation and was observed also for some other sequences of tolerances. The magnitude of the error can be computed from (10) assuming that $|b_{21}|_{L_1}, |f|_{L_1} = n\tau$ where n is the number of steps. This estimate is directly related to the fact that "per step" strategy leads to the optimal distribution of steps minimizing the perturbation in the L_1-norm. See [7].

The methods (M4), (M5), (M6) show similar performance although method (M5) gives better results than (M4), (M6). The factor common to (M4) and (M6) is the relation of u and w_2 with a negative sign in the differential equation (for $s \approx 0$ where the boundary layer is located) while (M5) uses the relation with opposite sign. The different character of perturbations then affects the error in $u'(0)$. We see that the transformation of the original problem into the first system and the choice of norm affects the error in $u'(0)$ because it leads to different characterizations of the discretization errorsas perturbations of the input data.

Table 4.1

Method 1. Error in $u'(0)$ and number of steps backward.

	absolute tolerance							
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	3.0820E-06 1	3.0820E-06 1	3.0820E-06 1	1.0248E-07 2	3.9468E-08 3	1.2039E-08 5	1.7953E-09 7	3.0876E-10 11
1.0000E-01	1.5190E-04 3	1.5190E-04 3	1.7454E-05 4	2.0155E-05 6	1.7975E-06 8	1.0358E-07 13	1.2759E-08 20	1.1353E-10 31
1.0000E-02	1.1043E-03 4	9.7234E-06 5	5.4961E-06 5	2.1295E-06 6	6.0009E-08 8	3.1869E-09 11	2.0307E-09 17	1.8775E-10 25
1.0000E-03	1.8641E-02 6	1.8641E-02 6	3.4241E-03 6	3.7770E-05 7	2.0772E-05 8	2.2579E-07 10	9.4277E-08 13	6.3624E-08 18
1.0000E-04	2.3018E-01 7	2.3018E-01 7	2.3018E-01 7	3.5308E-02 8	1.9995E-03 8	7.3388E-04 9	1.7008E-05 11	1.1898E-06 14
1.0000E-05	2.2905E+00 9	2.2905E+00 9	2.2905E+00 9	2.2905E+00 9	3.8971E-01 9	2.0070E-02 10	7.3669E-03 11	1.7051E-04 13
1.0000E-06	2.2825E+01 11	2.2825E+01 11	2.2825E+01 11	2.2825E+01 11	2.2825E+01 11	3.8986E+00 11	2.0208E-01 11	7.4113E-02 12
1.0000E-07	2.2941E+02 12	2.2941E+02 12	2.2941E+02 12	2.2941E+02 12	2.2941E+02 12	2.2941E+02 12	3.9009E+01 12	2.0208E+00 13

Table 4.2

Method 2. Error in u'(0) and number of steps backward.

				absolute tolerance				
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	3.0820E-06	3.0820E-06	3.0820E-06	1.0248E-07	3.9468E-08	1.2039E-08	1.7953E-09	3.0876E-10
	1	1	1	2	3	5	7	11
1.0000E-01	1.5190E-04	1.7454E-05	2.0155E-05	1.7975E-06	1.0358E-07	1.2759E-08	1.1352E-10	1.6356E-10
	3	4	6	8	13	20	31	50
1.0000E-02	5.4961E-06	2.1295E-06	6.0009E-08	3.1869E-09	2.0308E-09	1.8774E-10	2.1316E-13	3.3111E-12
	5	6	8	11	17	25	38	59
1.0000E-03	3.7770E-05	2.0772E-05	2.2579E-07	9.4278E-08	6.3624E-08	6.3883E-09	3.4106E-12	1.8645E-11
	7	8	10	13	18	26	39	61
1.0000E-04	1.9995E-03	7.3388E-04	1.7008E-05	1.1898E-06	6.2631E-07	6.3024E-08	3.8199E-11	6.8394E-10
	8	9	11	14	20	28	41	62
1.0000E-05	2.0070E-02	7.3669E-03	1.7051E-04	1.1849E-05	6.4008E-06	6.4434E-07	3.0559E-10	6.9849E-09
	10	11	13	16	21	29	43	64
1.0000E-06	2.0208E-01	7.4113E-02	1.7221E-03	1.1870E-04	6.3983E-05	6.4401E-06	3.1432E-09	7.0315E-08
	11	12	14	17	23	31	44	65
1.0000E-07	2.0208E+00	7.4116E-01	1.7222E-02	1.1869E-03	6.3994E-04	6.4436E-05	2.7940E-08	6.8545E-07
	13	14	16	19	24	32	46	67

Table 4.3

Method 3. Error in u'(0) and number of steps backward.

				absolute tolerance				
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	1.1738E-05	1.1738E-05	1.0766E-07	4.1489E-08	8.6509E-09	4.2592E-09	6.7639E-10	1.2762E-10
	1	1	2	4	5	6	9	14
1.0000E-01	1.5190E-04	1.5190E-04	1.7454E-05	2.0155E-05	1.7975E-06	1.0358E-07	1.2759E-08	1.1356E-10
	3	3	4	6	8	13	20	31
1.0000E-02	1.1043E-03	9.7234E-06	5.4961E-06	2.1295E-06	6.0009E-08	3.1868E-09	2.0307E-09	1.8764E-10
	4	5	5	6	8	11	17	25
1.0000E-03	1.8641E-02	1.8641E-02	3.4241E-03	3.7770E-05	2.0772E-05	2.2579E-07	9.4278E-08	6.3624E-08
	6	6	6	7	8	10	13	18
1.0000E-04	2.3018E-01	2.3018E-01	2.3018E-01	3.5308E-02	1.9995E-03	7.3388E-04	1.7008E-05	1.1898E-06
	7	7	7	8	8	9	11	14
1.0000E-05	2.2905E+00	2.2905E+00	2.2905E+00	2.2905E+00	3.8971E-01	2.0070E-02	7.3669E-03	1.7051E-04
	9	9	9	9	9	10	11	13
1.0000E-06	2.2825E+01	2.2825E+01	2.2825E+01	2.2825E+01	2.2825E+01	3.8986E+00	2.0208E-01	7.4113E-02
	11	11	11	11	11	11	11	12
1.0000E-07	2.2941E+02	2.2941E+02	2.2941E+02	2.2941E+02	2.2941E+02	2.2941E+02	3.9009E+01	2.0208E+00
	12	12	12	12	12	12	12	13

Table 4.4

Method 4. Error in u'(0) and number of steps backward.

				absolute tolerance				
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	2.1082E-06	2.1082E-06	2.1082E-06	8.1426E-08	4.1899E-08	1.3130E-08	2.4948E-09	4.1136E-10
	1	1	1	2	3	4	7	11
1.0000E-01	1.4772E-04	2.3294E-04	4.9107E-05	8.4943E-06	4.4914E-07	5.3726E-08	4.6513E-09	1.7075E-10
	3	3	5	7	10	16	25	39
1.0000E-02	5.5036E-05	4.9424E-09	7.6923E-07	8.2960E-08	1.1879E-09	8.3691E-10	3.0013E-11	3.3396E-12
	5	5	7	9	13	20	30	47
1.0000E-03	2.0253E-03	8.9701E-05	2.2672E-05	1.7898E-06	3.6016E-08	2.8657E-09	4.4145E-10	6.1846E-11
	6	7	8	11	15	22	32	49
1.0000E-04	2.0639E-02	1.1727E-03	2.2109E-04	2.7409E-05	3.6452E-07	1.9823E-07	7.1304E-10	1.5270E-08
	8	8	10	12	17	23	33	50
1.0000E-05	2.1325E-01	1.1697E-02	2.2562E-03	2.7437E-04	2.4974E-06	1.6795E-06	7.3608E-07	1.8963E-07
	9	10	11	14	18	25	35	52
1.0000E-06	2.1326E+00	1.1732E-01	2.2541E-02	2.5884E-03	1.2741E-04	7.3228E-05	1.5972E-05	2.1769E-05
	11	11	13	15	20	26	36	53
1.0000E-07	2.1335E+01	1.1833E+00	2.1307E-01	2.6545E-02	2.8526E-03	8.6018E-04	5.2899E-07	2.9476E-03
	12	13	15	17	21	28	38	55

Table 4.5

Method 5. Error in u'(0) and number of steps backward.

				absolute tolerance				
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	1.1738E-05	1.1738E-05	1.0766E-07	4.1489E-08	8.6509E-09	4.2592E-09	6.7639E-10	1.2762E-10
	1	1	2	4	5	6	9	14
1.0000E-01	8.5646E-05	7.5545E-05	2.8118E-05	2.3343E-06	1.1942E-07	9.6284E-09	9.8280E-11	1.5913E-10
	5	5	7	11	15	22	34	54
1.0000E-02	1.1686E-05	3.0201E-06	5.2790E-08	1.0821E-09	4.3633E-10	3.8582E-11	6.6507E-12	1.4211E-14
	7	7	10	14	19	27	41	64
1.0000E-03	4.6703E-04	4.0119E-05	1.5490E-06	4.7685E-08	1.2338E-09	9.9431E-10	1.7144E-10	4.4793E-11
	8	9	11	16	21	28	42	65
1.0000E-04	4.7695E-03	3.8288E-04	1.4605E-05	5.7743E-07	1.0795E-07	9.3514E-09	1.6225E-09	4.5657E-10
	10	11	13	17	22	30	44	67
1.0000E-05	4.9322E-02	4.6868E-03	1.4541E-04	5.7367E-06	1.0817E-06	9.3336E-08	1.6196E-08	4.6712E-09
	11	12	15	19	24	32	46	68
1.0000E-06	4.9322E-01	4.6883E-02	1.4820E-03	5.7360E-05	1.0935E-05	9.4832E-07	1.6438E-07	4.6915E-08
	13	14	16	21	25	33	47	70
1.0000E-07	4.9335E+00	4.6923E-01	1.4820E-02	5.7447E-04	1.0935E-04	9.4790E-06	1.6540E-06	4.7311E-07
	14	15	18	22	27	35	49	71

Table 4.6

Method 6. Error in u'(0) and number of steps backward.

	absolute tolerance							
	1.0000E-01	1.0000E-02	1.0000E-03	1.0000E-04	1.0000E-05	1.0000E-06	1.0000E-07	1.0000E-08
epsilon								
1.0000E+00	6.2792E-05	6.2792E-05	3.6794E-06	1.1308E-08	3.2130E-08	6.6043E-09	1.4241E-09	2.3792E-10
	1	1	2	3	5	7	11	17
1.0000E-01	2.0478E-04	2.0604E-04	2.2192E-05	2.1450E-06	1.7908E-07	5.2319E-08	2.2538E-09	4.2261E-11
	3	5	6	10	15	23	36	57
1.0000E-02	2.1116E-05	1.2709E-05	2.7052E-07	6.1835E-08	1.3435E-09	6.4652E-10	7.4323E-12	9.8055E-12
	6	7	9	13	19	27	42	65
1.0000E-03	2.9586E-04	6.2853E-05	2.4616E-05	3.0131E-07	4.6717E-08	1.5789E-08	8.5493E-11	7.4124E-11
	8	8	10	14	20	28	43	65
1.0000E-04	3.7127E-03	6.1580E-04	2.4863E-04	2.6167E-06	7.0509E-07	1.7281E-07	3.2232E-09	1.3843E-09
	9	10	12	16	21	30	44	67
1.0000E-05	3.6951E-02	5.6695E-03	2.5589E-03	3.5099E-05	7.1716E-06	2.0176E-06	1.2621E-07	2.0875E-07
	11	12	13	17	23	31	46	68
1.0000E-06	3.6824E-01	6.2807E-02	2.5587E-02	3.4538E-04	7.8466E-05	1.1818E-05	1.3631E-05	1.0675E-05
	13	13	15	19	24	33	47	70
1.0000E-07	3.7011E+00	6.2879E-01	2.5754E-01	2.1143E-03	5.4721E-04	8.2971E-04	1.3391E-03	1.5071E-0
	14	15	16	20	26	34	49	71

4.2 An unstable singular perturbation problem

Consider the singular perturbation problem of turning point type

(1a) $$\varepsilon w'' + \sigma s w' = 0, \qquad -a < s < b,$$

(1b) $$w(-a) = 1, \qquad w(b) = 2.$$

For $\sigma = 1$ the problem is stable and has an interior layer (shock) at $s = 0$ (Fig. 4.1) for all $a,b > 0$. The solutions shown in Fig. 4.1 were computed with the tolerance $\tau = 10^{-4}$ for all shown ε.

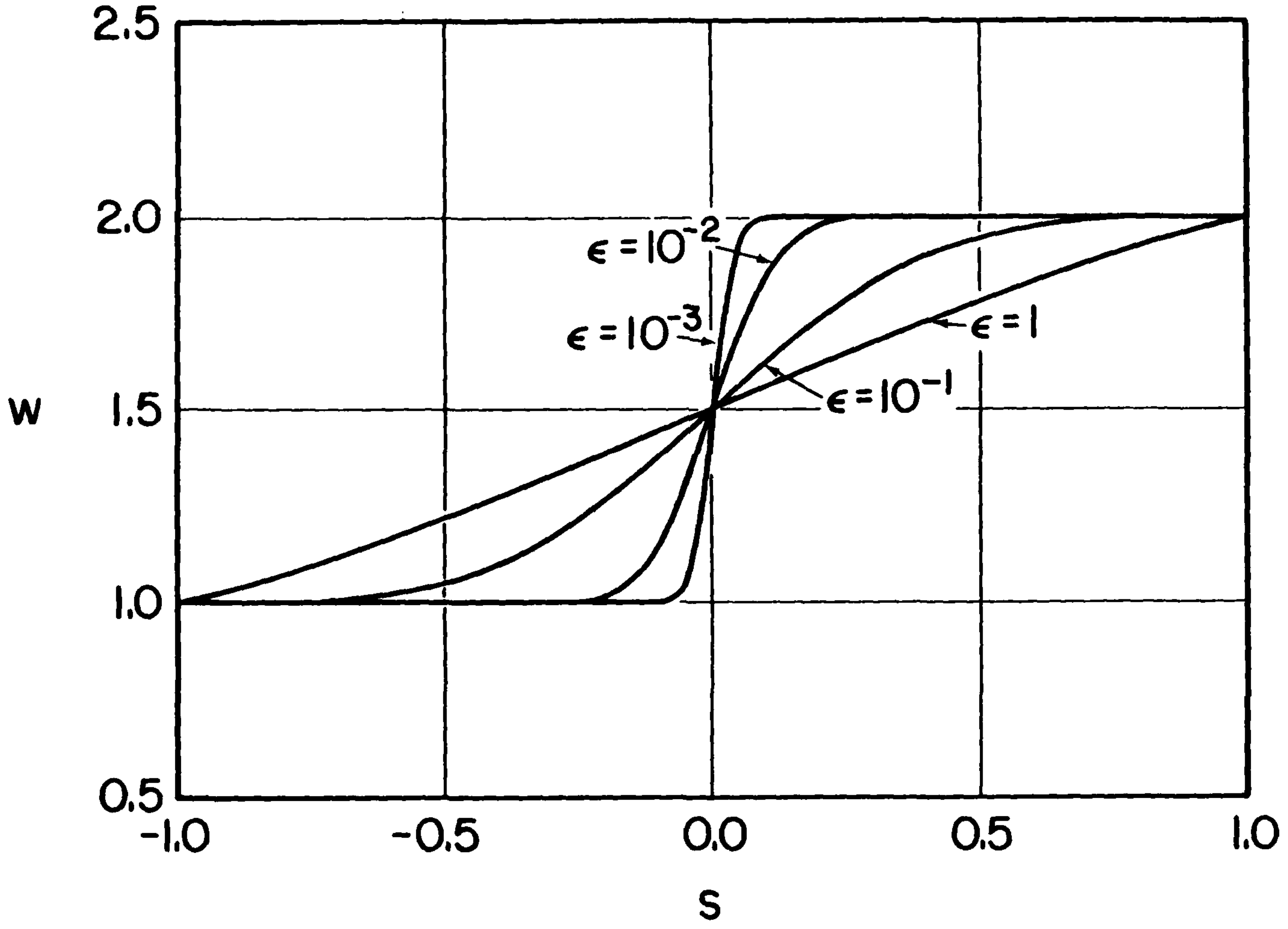


Fig. 4.1. Solution of the stable problem (la) (lb).

In the case $\sigma = -1$ a boundary layer occurs at $s = -a$ if $a > b$ and $s = b$ if $a < b$. If $a = b$, then there is a boundary layer at each end. The solution in this case $(a = b)$ is unstable with respect to the data as $\varepsilon \to 0$. The exact solution is antisymmetric with respect to the value 1.5. It can be computed for $s > 0$ by solving the (stable) problem

(2a) $$\varepsilon w'' - s w' = 0$$

(2b) $$w(0) = 1.5, \quad w(1) = 2.$$

Fig. 4.2 shows the solution for various ε computed by the factorization method with tolerance $\tau = 10^{-4}$. Solving the original problem (1) with $a = b = 1$, one has to expect that the results will be very poor for ε small because of the instability of the problem (1). Fig. 4.3 shows that in fact for ε small the factorization method completely fails. This example shows that the stability of the problem is a necessary condition for the factorization method to give high quality results. This condition is directly related to the interpretation of the numerical solution as the exact solution of a perturbed problem.

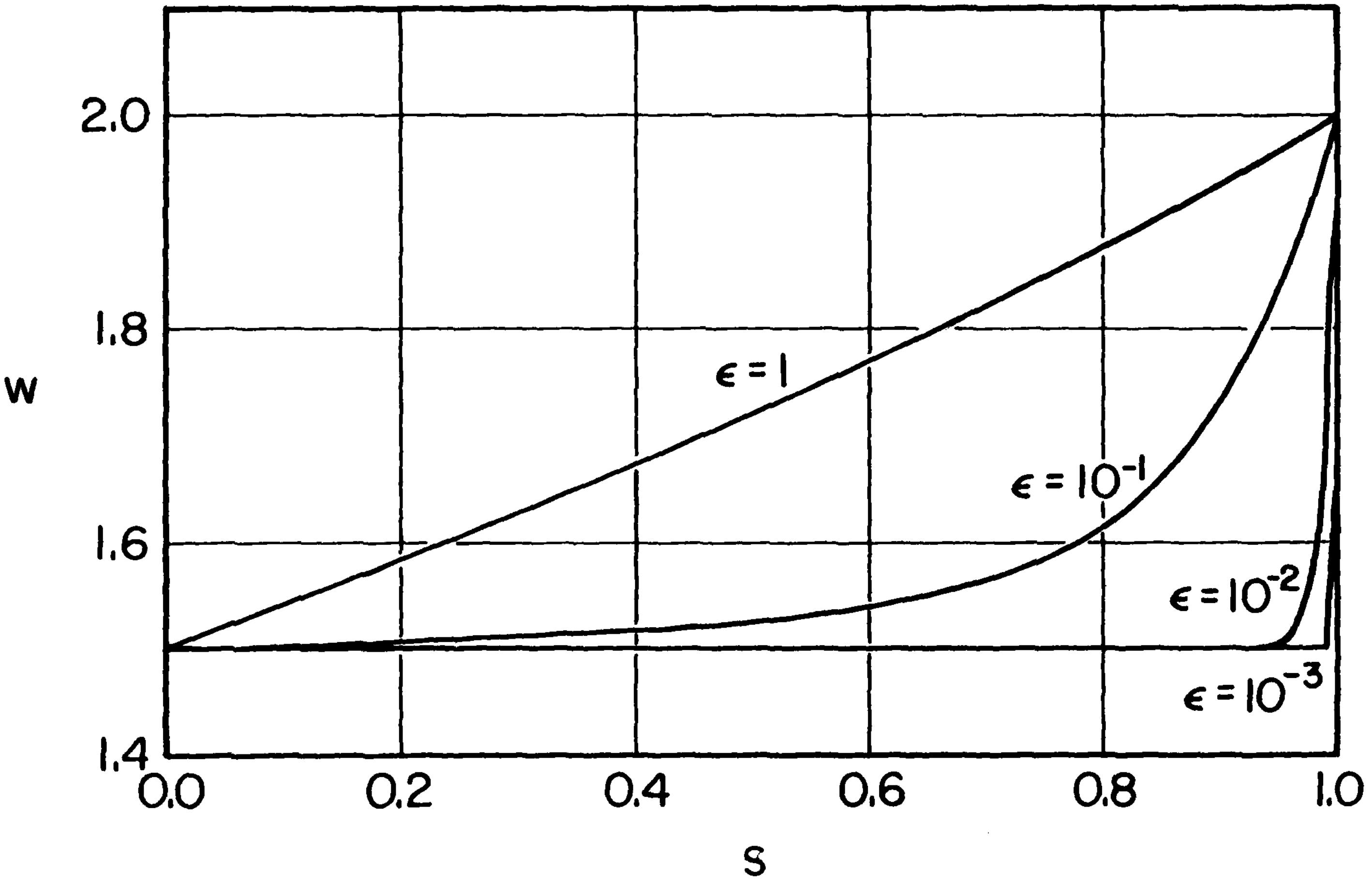


Fig. 4.2. Solution of the stable problem (2a) (2b).

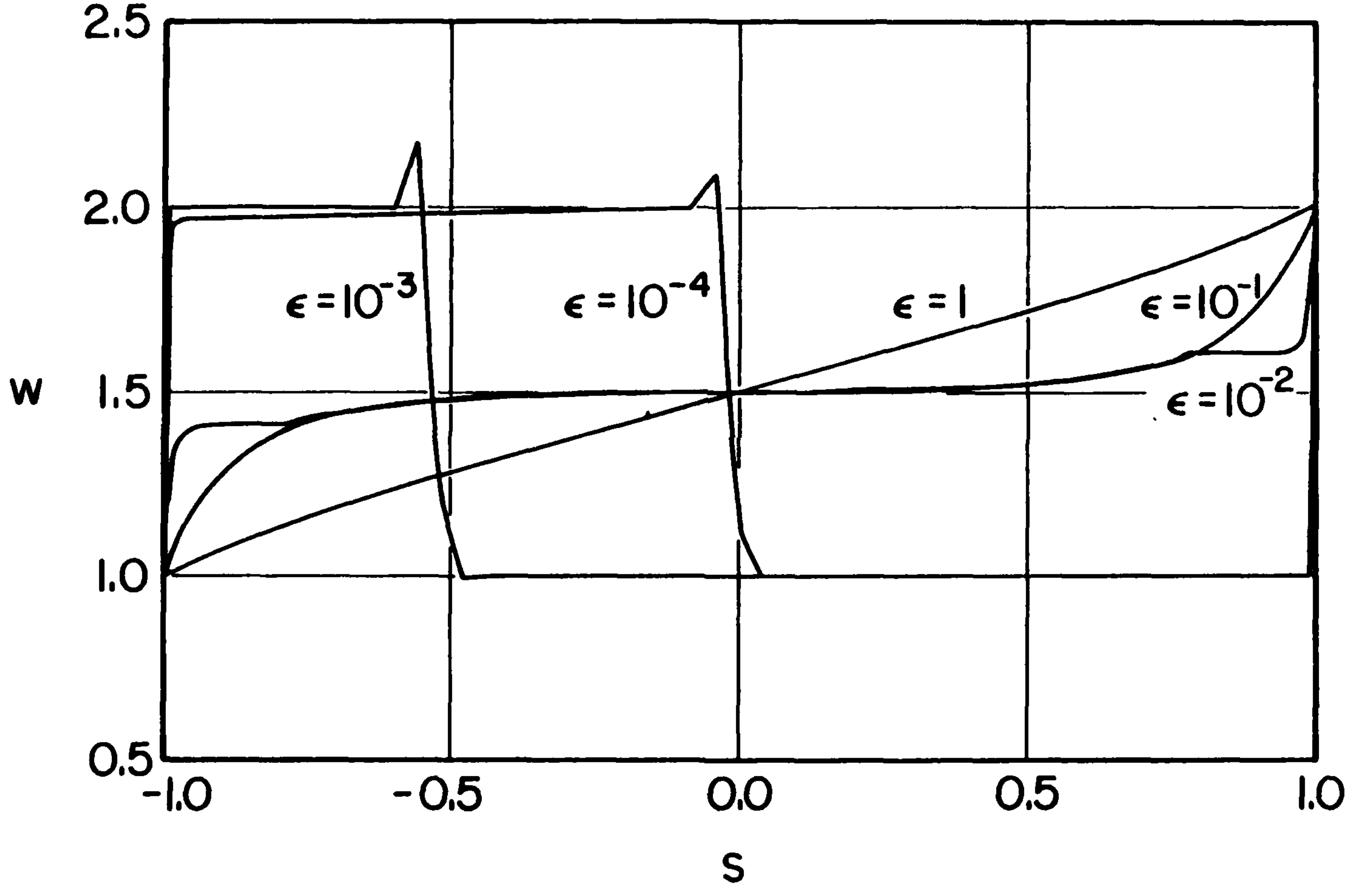


Fig. 4.3. Solution of the unstable problem (la) (lb) for
$\sigma = -1$, $a = b = 1$.

where L_k is the Legendre polynomial of degree k. Symmetry about $y = 0$ *in the* problem (1), (2) suggests an approximation for w of the form

$$w(x,y) = \sum_{j=0}^{N} w_j(x)\ell_{2j}(y) \tag{5}$$

for (x,y) Ω. The convergence and approximation properties of this method, called dimensional reduction, are discussed by Vogelius and Babuška [22]. The relevance of (5) for our purposes is that it reduces the PDE (1),(2) to a system of ODE's for the vector function

$$W(\cdot) \equiv (w_0(\cdot),w_1(\cdot),\ldots,w_N(\cdot))^T \tag{6}$$

mapping $[0,1]$ into $\mathbf{R}^{N+1}$. The function W is the solution of the two point boundary value problem

$$-hA\,\frac{d^2W(x)}{dx^2} + h^{-1}BW(x) = G(x), \qquad x \in (0,1) \tag{7a}$$

$$W(0) = W(1) = 0 \tag{7b}$$

where $G(x) = (2g(x),0,\ldots,0)^T$, and the $(N+1) \times (N+1)$ matrices A and B are the mass and stiffness matrices, respectively, of the basis $\{\ell_{2j}\}$. That is,

$$A_{ij} = \int_0^1 \ell_{2j}(y)\ \ell_{2j}(y)dy, \tag{8}$$

and

$$B_{ij} = \int_0^1 \frac{d\ell_{2i}}{dy}(y)\ \frac{d\ell_{2j}}{dy}(y)dy. \tag{9}$$

We note that the presence of the parameter h in (7a) gives the TPBVP (7) a singular perturbation character if $h << 1$.

Figures 4.4 and 4.5 are plots of the solution components $\{w_k\}_{k=0,\ldots,N}$ and $\{w_k'\}_{k=0,\ldots,N}$ for the case $N = 4$ and $h = 0.5$, with g given by

$$(10) \qquad g(x) = \begin{cases} 1, & 0.475 \le x \le 0.525 \\ 0, & \text{otherwise.} \end{cases}$$

Only the solution on $0 \le x \le 0.5$ is plotted since $w(x,y)$ in this case is symmetric about $x = 0.5$ and $y = 0$. Evidently the component w_0 dominates the solution except in the interior layer caused by the flux, g. A contour plot of w obtained from expansion (5) for $h = 0.5$ and $N = 4$ is shown in Figure 4.6. The singularity at the corner due to the step in the flux is evident.

The computations were done with tolerance of $\tau = 10^{-4}$. In view of (3.2-4), this amounts to computing the <u>exact</u> solution to the first order system

$$(11a) \qquad \frac{d}{dx} W_1(x) - [Id_N + b(x)]W_2(x) = f(x)$$

$$(11b) \qquad -hA \frac{d}{dx} W_2(x) + h^{-1}BW_1(x) = G(x)$$

$$(11c) \qquad W_1(0) = W_1(1) = 0$$

where $W_1 = W$ and the perturbing matrix $b(x)$ and vector $f(x)$ are unknown but satisfy

$$(12) \qquad |b_{ij}(x)|, \ |f_i(x)| < \tau, \quad i,j = 0,N, \quad x \in (0,1).$$

This in turn implies that the numerical solution W_1 is the exact solution for problem (1) in which the flux G is replaced by the flux

$$G(x) - h\Delta \frac{d}{dx} (bW_2(x) + f(x)).$$

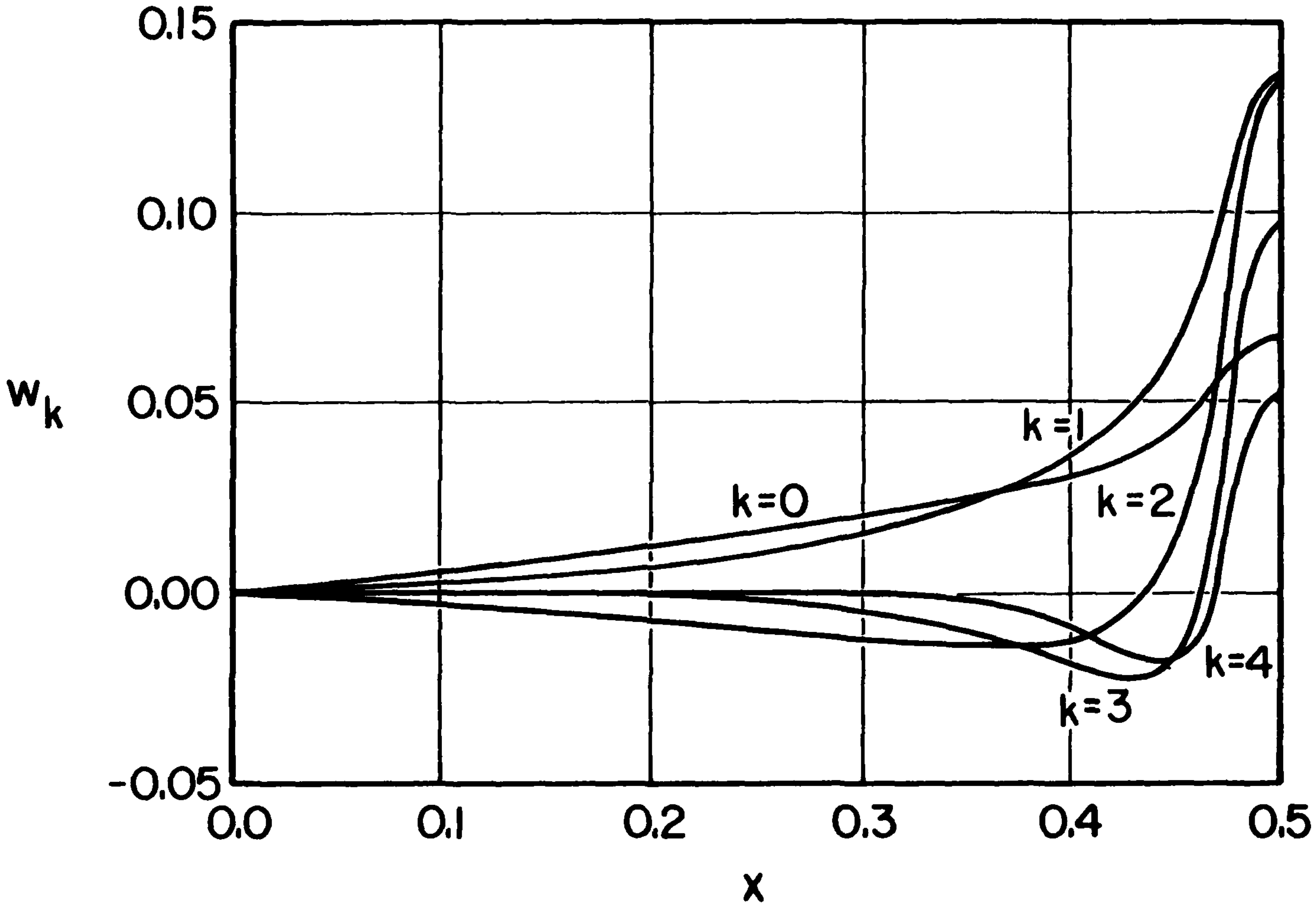


Fig. 4.4. Coefficient functions.

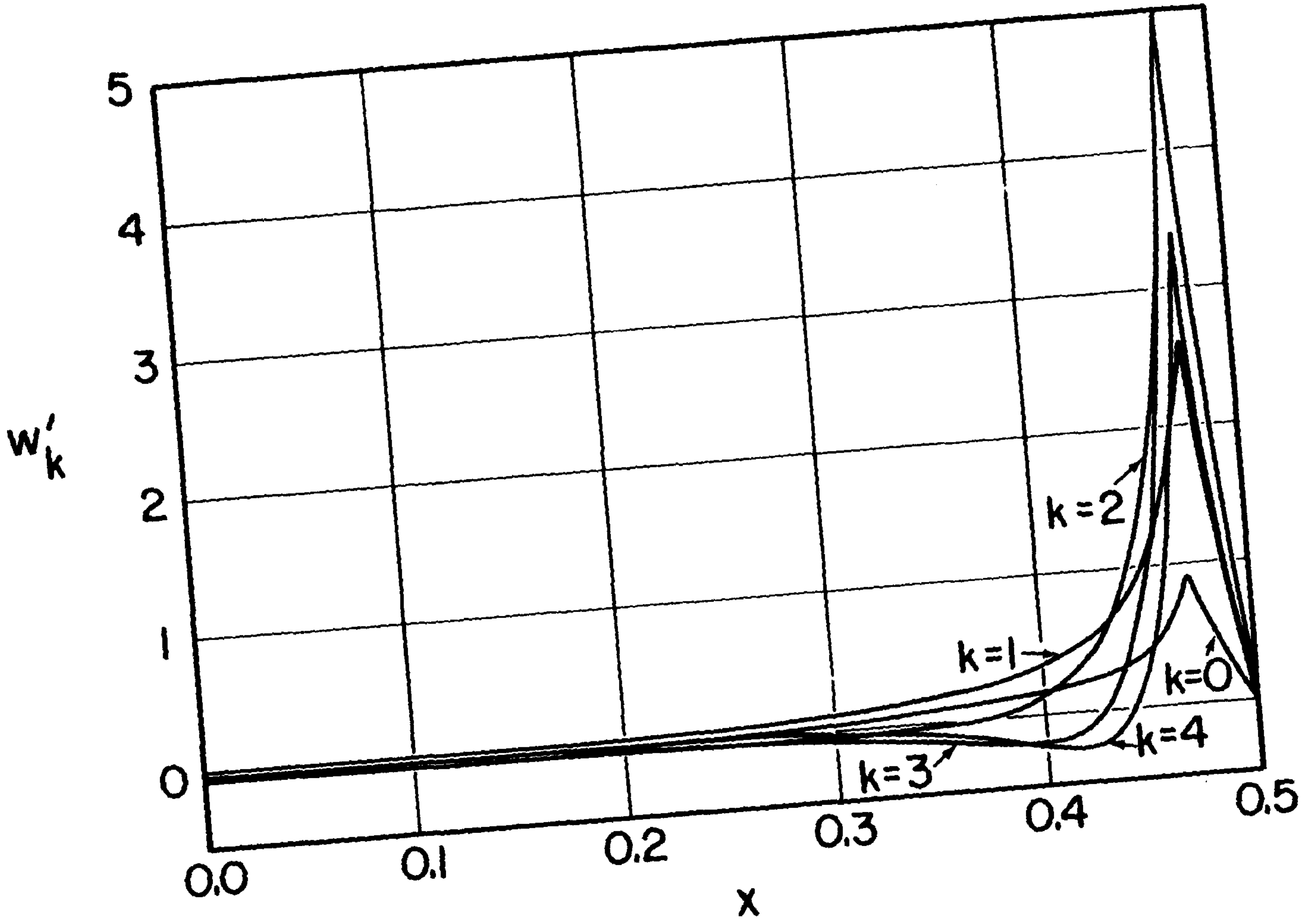


Fig. 4.5. Derivatives of coefficient functions.

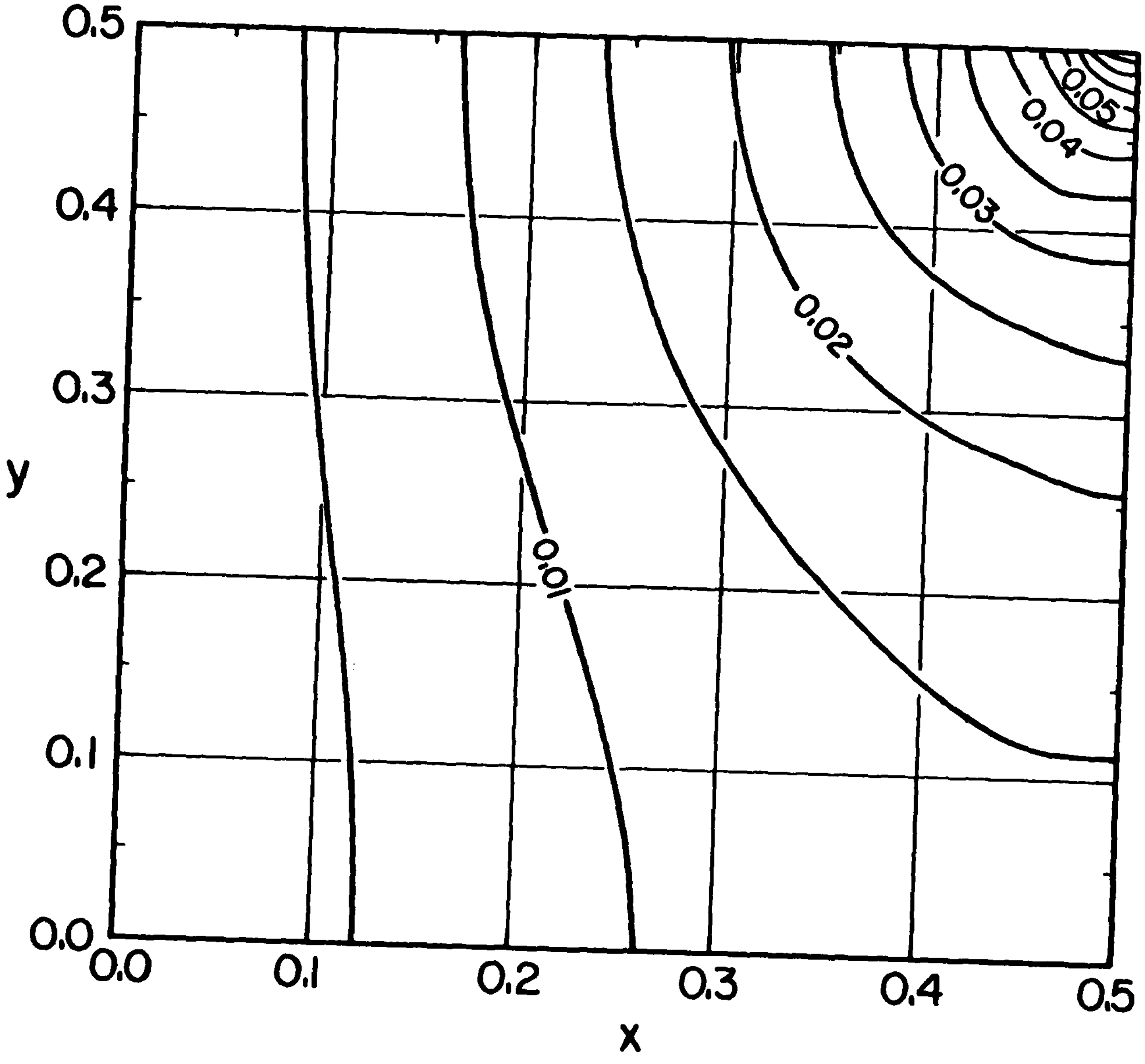


Fig. 4.6. Contours of w.

REFERENCES

[1] V. Ascher, J. Christiansen, and R.D. Russel, A collocation solver for mixed order systems of boundary value problems, Math. Comp. 33 (1979), 659-674.

[2] I. Babuška, The connection between finite difference-like methods and methods based on initial value problems for ODE's, in Numerical Solution of Boundary Value Problems for ODE's A. K. Aziz (ed.), Academic Press (1974).

[3] I. Babuška and V. Majer, The factorization method for two point boundary value problems for ODE's and its relation to the finite difference method, in Proceedings of the Centre for Mathematical Analysis of the Australian National University, Vol. 7, A. Miller (ed.) (1984), 71-92.

[4] I. Babuška, M. Práger, and E. Vitásek, Numerical Processes in Differential Equations, Wiley, London (1966).

[5] I. Babuška, M.Práger, and E. Vitásek, The closure of numerical processes and the method of factorization, Zurn. Vyc. Mat. i Mat. Fiz. (1964), 351-353.

[6] I. Babuška and S.L. Sobolev. The optimization of numerical processes, Apl. Mat. 10 (1965), 96-130.

[7] I. Babuška and M.S. Vogelius. Optimal error control strategies for initial value problems, to appear.

[8] R.E. Bellman, Introduction to Matrix Analyis, McGraw-Hill, New York (1960).

[9] K. Burrage and J.C. Butcher, Stability criteria for implicit Runge-Kutta methods, SIAM J. Num. Anal. 16:1, 46-57.

[10] A. Davey, An automatic orthonormalization method for solving stiff boundary value problems, J. Comp. Phys. 51 (1983), 343-356.

[11] C.G. Guderley, A unified view of some methods for stiff two point boundary value problems, SIAM Rev. 17 (1975), 416-442.

[12] P.W. Hemker, A Numerical Study of Stiff Two Point Boundary Value Problems, Mathematics Centrum, Amsterdam (1977).

[13] H.B. Keller and M. Lentini, Invariant imbedding, the box scheme, and an equivalence between them, SIAM J. Num. Anal. 19 (1982), 942-962.

[14] A.J. Lamb, A Schur method for solving algebraic Riccati equations, IEEE Trans. Aut. Control 24:6 (1979), 913-921.

[15] M. Lentini, M.R. Osborne, and R.D. Russell, The close relationship between methods for solving two point boundary value problems, SIAM J. Num. Anal.

[16] M. Lentini and V. Pereyra, An adaptive finite difference solver for nonlinear two point boundary value problems with mild boundary layers, SIAM J. Num. Anal. 14 (1977).

[17] R.M.M. Mattheij and G.W.M. Staarink, An efficient algorithm for solving general linear two point BVP, SIAM J. Sci. Stat. Comput. 5 (1984).

[18] R.M.M. Mattheij, Decoupling and stability of algorithms for boundary value problems, SIAM Review 27 (1985), 1-44.

[19] V. Majer, Numerical solution of boundary value problems for ordinary differential equations of nonlinear elasticity, Ph.D. Thesis, Univ. of Maryland, 1984.

[20] G.H. Meyer, Continuous orthonormalization for boundary value problems, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.

[21] J.E. Potter, Matrix quadratic solutions, SIAM J. Appl. Math. 14 (1966), 496-501.

[22] J. Taufer, On the factorization method, Apl. Mat. 11:6 (1966), 427-451.

[23] J. Taufer, Faktorisierungsmethode für ein Randwertproblem eines linearen System von Differentialgleichungen, Apl. Mat. 13:2 (1968), 191-202.

[24] M.S. Vogelius and I. Babuška, On a dimensional reduction method.

I: The optimal selection of basis functions, Math. Comp. 37:155 (1981), 31-46.

II: Some approximation-theoric results, Math. Comp. 37:155 (1981), 47,69.

III: A posteriori error estimation and an adaptive approach, Math. Comp. 37:156 (1981), 361-384.

The Laboratory for Numerical analysis is an integral part of the Institute for Physical Science and Technology of the University of Maryland, under the general administration of the Director, Institute for Physical Science and Technology. It has the following goals:

- To conduct research in the mathematical theory and computational implementation of numerical analysis and related topics, with emphasis on the numerical treatment of linear and nonlinear differential equations and problems in linear and nonlinear algebra.

- To help bridge gaps between computational directions in engineering, physics, etc., and those in the mathematical community.

- To provide a limited consulting service in all areas of numerical mathematics to the University as a whole, and also to government agencies and industries in the State of Maryland and the Washington Metropolitan area.

- To assist with the education of numerical analysts, especially at the postdoctoral level, in conjunction with the Interdisciplinary Applied Mathematics Program and the programs of the Mathematics and Computer Science Departments. This includes active collaboration with government agencies such as the National Bureau of Standards.

- To be an international center of study and research for foreign students in numerical mathematics who are supported by foreign governments or exchange agencies (Fulbright, etc.)

Further information may be obtained from Professor I. Babuska, Chairman, Laboratory for Numerical Analysis, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

END

FILMED

3-86

DTIC